Using Georeferenced Data in Social Science Survey Research: The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel

Jünger, Stefan

gesis Leibniz-Institut für Sozialwissenschaften

Schriftenreihe

Band 24

Stefan Jünger

Using Georeferenced Data in Social Science Survey Research

The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel

Using Georeferenced Data in Social Science Survey Research

GESIS Series
published by GESIS – Leibniz Institute for the Social Sciences

Volume 24

Stefan Jünger

Using Georeferenced Data in Social Science Survey Research.
The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel

Die vorliegende Arbeit wurde am Institut für Soziologie und Sozialpsychologie (ISS) der Wirtschafts- und Sozialwissenschaftlichen Fakultät der Universität zu Köln als Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschafts- und Sozialwissenschaften (Dr. rer. pol.) angenommen.

Stefan Jünger

Using Georeferenced Data in Social Science Survey Research

The Method of Spatial Linking and Its Application with the German General Social Survey and the GESIS Panel

GESIS – Leibniz Institute for the Social Sciences 2019

Bibliographical information of the German National Library (DNB)

The German National Library lists this publication in the German National Bibliography; detailed bibliographical data are available via https://www.dnb.de.

ISBN	978-3-86819-040-3 (print)
ISBN	978-3-86819-039-7 (eBook)
ISSN	1869-2869

Publisher, printing and distribution:

GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln, Tel.: 0221 / 476 94 - 0
publications@gesis.org
Printed in Germany

Contents

Abbreviations & Acronyms

AIC	Akaike Information Criterion
API	Application Programming Interface
ANNOY	Noise Annoyance
BIC	Bayesian Information Criterion
BKG	Federal Agency for Cartography and Geodesy
CFI	Comparative Fit Index
CI	Confidence Interval
CRS	Coordinate Reference System
CSV	Comma Separated Values
dB(A)	A-weigthed decibels (low frequencies are reduced)
DE-9IM	Dimensionally Extended Nine-Intersection Model
destatis	Federal Statistical Office of Germany
DSAnpUG-EU	Law for Adaptation of Data Protection to the Regulation of the European Union
EIONET	European Environment Information and Observation Network
EPSG	European Petroleum Survey Group Geodesy
EU	European Union
EUR	Euro, currency
FE	Fixed Effects
GDPR	General Data Protection Regulation
GeorefUm	Georeferencing of Survey Data Project
GGSS	German General Social Survey
GIS	Geographic Information Systems
GLES	German Longitudinal Election Study
GPS	Global Positioning System
HRQL	Health-related Quality of Life
IOER	Leibniz Institute for Urban and Ecological Development
ISCED	International Standard Classification of Education

IV	Instrumental Variable
km	kilometer
m	meter
MAUP	Modifiable Area Unit Problem
MCS	Mental Component Score
ML	Multilevel (Models)
MLR	Maximum Likelihood with Robust Standard Errors
NEPS	German National Educational Panel Study
OGC	Open Geospatial Consortium
OLS	Ordinary Least Squares
OSM	OpenStreetMap
OVB	Omitted Variable Bias
PCA	Principal Component Analysis
PCS	Physical Component Score
pairfam	Panel Analysis of Intimate Relationships and Family Dynamics
RMSEA	Root Mean Square Error of Approximation
RTN	Road Traffic Noise
SD	Standard Deviation
SE	Standard Error
SEM	Structural Equation Model
SOEP	Socio-economic Panel
SoRa	Social-Geospatial Data Infrastructure Project
TLI	Tucker-Lewis Index
UBA	German Environmental Agency
UK	United Kingdom
URL	Uniform Resource Locator
US(A)	United States (of America)
WCS	Web Coverage Service

Figures

Tables

1 Introduction

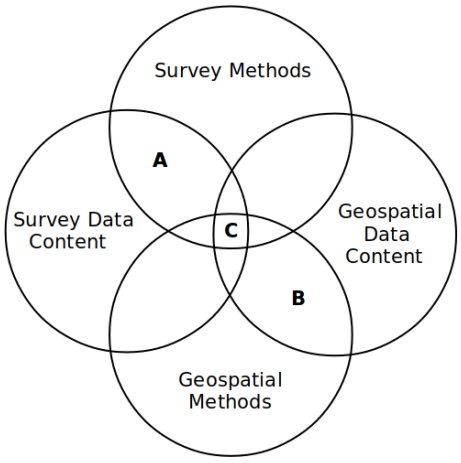
Georeferenced social science survey data are survey data enriched with direct spatial identifiers, such as geo-coordinates (Meyer & Bruderer Enzler, 2013, 323). Researchers use them to locate survey respondents in geographic space, to integrate spatial relationships in their analyses, and to add geospatial information from other data sources to their survey data (Chhetri & Stimson, 2014), which defines the method of spatial linking. This effort, declared as part of the "spatially integrated social science" (Goodchild, Anselin, Appelbaum, & Harthorn, 2000), in recent years in some disciplines even labeled as "spatial turn" (Richardson et al., 2013), gained a lot of popularity in current research applications (for a general overview see Bluemke, Resch, Lechner, Westerholt, & Kolb, 2017, and for the particular analysis potential of georeferenced survey data Hillmert, Hartung, & Weßling, 2017).

Georeferenced survey data are advertised to increase the knowledge about social phenomena, for example, by providing "new vehicles for innovation, synthesis and integration across the social and behavioral sciences" (Stimson, 2014, 13). Thus, by locating survey respondents in space, researchers gain unprecedented insights into the contexts of social behavior and attitudes. They can include characteristics of the respondents' direct living environment on different geographic levels, e.g., streets, building blocks, city districts, or entire municipalities. Accordingly, applications can be found in an extensive collection of social sciences' sub-disciplines ranging from political behavior and attitudes (Dill & Jirjahn, 2014; Förster, 2018; Klinger, Müller*, & Schaeffer, 2017), to educational (Ainsworth, 2002; Crowder, Hall, & Tolnay, 2011; Weßling, 2016) and health research (Bocquier et al., 2013; Oiamo, Baxter, Grgicak-Mannion, Xu, & Luginaah, 2015; Saib et al., 2014).

However, by using georeferenced survey data social scientists face challenges, starting with the question of how to conceptualize respondents' direct living environment, which is the neighborhood (Dietz, 2002; Foster & Hipp, 2011; Kwan, 2012; Perchoux, Chaix, Brondeel, & Kestens, 2016; Sampson, Morenoff, & Gannon-Rowley, 2002; Spielman & Yoo, 2009). As some of the geospatial information is available on a particular small scale, researchers have to decide what amount of information they want to include in their data and analysis (Wu, 2007, 199f). They have to weigh between considering geographic information in a range of, e.g., 1000 meters around survey respondents' locations for neighborhood operationalizations, or to use information on a broader range. Often, social science theories do not provide any help to this decision. For this reason, recent studies frequently examine variations of neighborhood indicators, depending on the geographic scale of the included information (Klinger et

al., 2017; Sluiter, Tolsma, & Scheepers, 2015; Tolsma & van der Meer, 2017).

Moreover, challenges also occur after adding geospatial information from auxiliary data sources to georeferenced survey data. As this information often stems from other scientific disciplines, such as ecology, engeneering, or the spatial sciences, social science research turns to an interdisciplinary undertaking (Dietz, 2002, 540). For example, if social scientists want to study the social conditions of road traffic noise and health (Passchier-Vermeer & Passchier, 2000), a potential data source is the already available road traffic noise data measured as sound pressure levels on a decibel value scale (European Parliament & European Council, 2002, Annex I). In contrast to established and validated measures on attitudes or behaviors in standardized surveys, these measures are foreign to survey researchers. The data generation process differs from interviewing in survey research and depends on the discipline from which geospatial information is added. While survey data's units of observation are people who are interviewed, geospatial data's are geometries that can be plotted on a map. With regards to environmental noise data, for example, social scientists have to make themselves familiar with the methods and data of acoustic engineering. After acquiring and preparing data from other sources (Schweers, Kinder-Kurlanda, Müller*, & Siegers, 2016, 107ff), using georeferenced survey data remains an effort that requires practice in the methods of entirely different scientific disciplines.



Note: **A** corresponds to the combination of expertise in the social sciences; **B** to the expertise in scientific disciplines that produce geospatial data; and **C** to the combination between **A** and **B** required to conduct research with georeferenced survey data

Figure 1.1: Schematic Display of Interdisciplinarity in Research with Georeferenced Survey Data

Figure 1.1 displays a scheme of the interdisciplinarity that is involved in conducting

research with georeferenced survey data. Social scientists that work with survey data are experts in the method of survey research and the actual content of these data (**A**). Likewise, researchers in other disciplines who produce geospatial data are experts in corresponding geospatial methods and the content of the geospatial data (**B**). By using survey data in combination with geospatial data, social scientists have to be familiar with all four dimensions of expertise: survey methods, survey data content, geospatial methods, and geospatial data content (**C**). Acquiring this expertise is challenging and gripping, which was the motivation to write this dissertation.

The guiding question of this dissertation, thus, relates to the above-sketched problem areas: *What is the actual gain in knowledge of using georeferenced survey data in the social sciences that makes addressing these challenges worthwhile*? By disentangling the challenges of the spatial data landscape in Germany from the general strengths and promises of georeferenced survey data, answers to the following more specific questions are provided:

- How can researchers add geospatial information to social science survey data?
- What data and methods are available for spatial linking?
- Which social science research applications benefit from georeferenced survey data?
- To conduct promising research, what criteria can researchers use to choose from the data and methods described earlier?

The particular emphasis of this analysis is on the theoretical and methodological implications and consequences of the spatial linking method, i.e., the enrichment of survey respondents' geo-coordinates with additional geospatial information. Accordingly, this work provides an interface between the theoretical conceptualization of neighborhoods (Spielman & Yoo, 2009) and the actual statistical analysis of the resulting data (Hillmert et al., 2017). While also addressing both of these topics, this work investigates the methodological challenges of adding geospatial information to georeferenced survey data.

1.1 Main Findings

This book provides examples and applications for social science research which use georeferenced survey data in Germany. It reasons the scientific prerequisites and introduces the challenges of this effort. In the following, it navigates through applica-

tions with different research questions, georeferenced data, and geospatial data. Accordingly, the results of these applications lead to different conclusions. This diversity of conclusions offers a realistic picture of the research landscape with georeferenced survey data in Germany. The following points summarize the main findings of this book:

1. *Effects in statistical models are often small.* The promise of unprecedented insights into the contexts of social behavior and attitudes by using georeferenced survey data sounds auspicious. This statement, though, does not necessarily mean that magnitudes of effects in statistical models are more pronounced or statistically more significant. The application of immigrant rates in small-scale neighborhoods and their effect on xenophobic attitudes in this book (Application II, Chapter 7) even yield null results. These findings are in stark contrast to previous research which used administrative districts data and found evidence for such effects. Seeking the most significant effects in statistical models may not be the ideal motivation to conduct research based on georeferenced survey data.

2. *Censored geospatial data lead to small sample sizes.* Most of the available sources for geospatial data were not collected in the context of social science research. For example, public authorities collected the road traffic noise data that are used in Application I of this book (Chapter 6) to create maps of noise exposure for whole neighborhoods in cities. Small roads that were not often frequented by cars were not part of the data collection. While it is still possible to use the data to gather information on road traffic noise for points on a map, the data are censored—noise values below some specific threshold are not part of these data. Population survey samples aim to collect data on the whole range of a population, ranging from people who live in small towns to those living in large cities. Moreover, a large proportion of these people's houses are located on small roads. Consequently, the combination of survey data and road traffic noise data results in small sample sizes of just ~10% of respondents with an actual road traffic noise measurement. This circumstance requires extra efforts to create valid estimates, which this application, for example, solves with an instrumental variable approach.

3. *Being close to respondents may not require complicated statistical models.* Using small-scale geospatial information has advantages for statistical models. Often, survey research that links auxiliary contextual information on large-scale regional units suffers from clustered observations in these regional units. Researchers use elaborated models such as multilevel models to account for this clustering and to aim for unbiased estimates. With contextual information from small-scale regional

units, this effort may not be necessary. While the survey data is still clustered because of the sampling procedures, adding geospatial information to the survey data does not necessarily create more dependencies between observations. Application III (Chapter 8) exemplifies this approach by using clustered standard errors for the sample, but it did not require to account for clustering of the geospatial data, in this case, on soil sealing.

4. *Chance to answer innovative research questions.* Researchers have been using methods of adding auxiliary data to survey data already for a long time. Often these data comprise auxiliary sociodemographic data on large-scale spatial units, operationalized as the societal or neighborhood context of a person. Small-scale geospatial data, on the other hand, are also available for information from different disciplines, e.g., ecology or the geospatial sciences. In combination with survey data, these data enable researchers to answer innovative research questions with indicators which have not been used before. Application I in this book (Chapter 6), for example, reassesses research questions from stress, family and health research with road traffic noise data, and Application III (Chapter 8) exploits rare land use indicators to operationalize environmental hazards exposure.

5. *Geospatial data enable more sophisticated operationalizations of neighborhoods.* Geospatial data enable social scientists to answer innovative research questions, and primarily small-scale geospatial data allow them to create measures of neighborhoods that are more elaborated. For example, as in Application II (Chapter 7), researchers can classify neighborhoods into direct neighborhoods and surrounding neighborhoods of people and relate these two types to each other. Or, as in Application I (Chapter 6), researchers calculate geodesic distances from survey respondents' locations to other points in the geospatial data. These methods create opportunities for research that differ severely from adding auxiliary data for large geographic units to survey data.

1.2 Organization of the Book

Chapter 2 starts with the basic terms and the general concepts for the use of georeferenced survey data. It introduces the methods of georeferencing and geocoding as well as the two data types geospatial data and georeferenced survey data. Furthermore, it provides a first discussion of the general method of spatial linking which is continued in Chapter 5. Chapter 2 concludes with a first look at the actual product of combining georeferenced survey data with geospatial data: information on spaces, places, and

neighborhoods as well as their relevance to social science theory.

Chapter 3 continues with examples for the use of georeferenced data in social science research. These examples include work from the research fields comprising the empirical applications in Chapter 6-8: health, political opinions, and social inequalities. The chapter closes with a discussion of the common methodological themes that occur within these fields.

Chapter 4 provides an overview of the numerous challenges of using georeferenced survey data. They involve challenges such as the availability of data, the application of technical procedures, and the legal barriers of survey respondents' data protection and privacy. This chapter does not only work up these considerations, but it also proposes solutions to these challenges.

Chapter 5 presents some of the numerous possibilities of adding geospatial information to georeferenced survey data. First, the difference to other approaches commonly used to link area information to survey data is explained. The following sections show different procedures of spatial linking, e.g., linking by location, buffers, focal neighborhoods and geodesic distances using Geographic Information Systems (GIS). The last step presents statistical models for analyzing the resulting data because using spatially clustered data can pose some severe statistical challenges in the analysis.

Chapter 6 comprises the first empirical application that builds on the previous elaborations by deploying them for family and health research. It combines data from the georeferenced German General Social Survey (GGSS) and geospatial data from road traffic noise to investigate whether married people report fewer health problems than unmarried people when they are exposed to road traffic noise stressors. The application's theory is deeply rooted in the study of social ties and health which is linked to existing research on environmental stressors. Structural equation models reveal a complex nexus of relationships between marriage, noise annoyance as well as mental and physical health. These models show indications that married people are less affected by exposure to road traffic noise stressors. The effects are smaller than other stress measures in the literature, but they give external validity to the theory that, in society, married people are still better off when it comes to stress, family and health.

Chapter 7 represents the second empirical application which exploits georeferenced survey data for research questions in the area of political attitudes. For this purpose, it links data from the georeferenced GGSS to geospatial data from the German Census 2011 and analyzes how ethnically diverse neighborhoods relate to the xenophobia of people that live in ethnically homogeneous neighborhoods. Given the evidence from other European countries, it is surprising that the analysis which uses dif-

ferent neighborhood operationalizations and estimation methods yields null-findings. These results illustrate how sensitive social science theories are to differences in societal contexts. In contrast to other countries, the segregation structure of Germany differs, and theories about such structures may not apply to the German context. The application of georeferenced survey data contributes to this discussion by providing detailed insights into the spatial integration of social science theories.

Chapter 8 which depicts the last empirical application uses georeferenced survey data for research in the area of environmental inequalities. Survey data from the georeferenced GESIS Panel are spatially linked to data about soil sealing from the Leibniz Institute for Urban and Regional Development (IOER) on a varying geographic scale. Furthermore, the application explores if people with a migration background are more often affected by soil sealing hazards and if income reduces their exposure. Competing theories about ethnic inequalities predict differential trajectories. As such, this application evaluates which of the theories is more likely to apply. While income reduces the risk of exposure for both the German and migrant group at a specific geographic level, some of the inequalities between these two groups remain even after the analysis. Thus, it is plausible that more than one theory may apply depending on specific income groups across German and migrant people. This application of georeferenced survey data also exemplifies the importance of particular small-scale geospatial data as the varying geographic scale discloses the sensitivity of the results with regards to the scale.

Chapter 9 is the conclusion of this book and summarizes all results. It reviews the interdisciplinary components of research with georeferenced survey data and discusses this effort as part of social science research that primarily emphasizes the context of social behavior. The aim is to assess the gain in knowledge through georeferenced survey data. The chapter ends with some notes on future avenues for social science research using other types of data, such as sensor or internet data.

2 Basic Terms, Data Types, and General Concepts

Using georeferenced survey data introduces new terms, data, and concepts to social science research. As an interdisciplinary effort at the interface between social sciences and geosciences, researchers require an understanding of the methods that create their applications' data. Moreover, exploiting these data has theoretical implications: in countless research applications, it is still unclear how and on which scale the geographic context impacts people's lives. Researchers have to clarify the basic terms, data types, and general concepts before they can conduct research with georeferenced survey data.

This chapter starts with an introduction to Geographic Information Systems (GIS) and defines the basic terms of georeferencing and geocoding. It moves on to the two data types of georeferenced survey data and geospatial data. Each introduction of these data types also presents the datasets which the empirical applications in Chapter 6-8 of this book use. The following section exhibits the combination of both data sources—the method of spatial linking (Chapter 5.2 discusses this method in more detail). A final theoretical discussion asks how all these efforts of involving and linking different data types can produce measurements for socially relevant contexts.

2.1 Geographic Information Systems (GIS)

To use geospatial data and methods of spatial analysis, researchers need access to and expertise in specialized software: Geographic Information Systems (GIS) (Bluemke et al., 2017). Exploiting GIS requires training because the data structure of geospatial data is uncommon for researchers with a background in survey research (Meyer & Bruderer Enzler, 2013, 319). Fortunately, in recent years, not only the amount of training offers increased, but also the market for GIS software evolved towards a more free and open source direction. Ordinary research projects no longer need to use commercial software, such as *ArcGIS* (ESRI, 2015), and can, instead, rely on open source software, such as *QGIS* (QGIS Development Team, 2019) or the statistical software *R* (R Core Team, 2019).[1] Moreover, GIS has developed to be a standard term in work with geospatial data. Scientific disciplines with an emphasis on spatial methods label themselves as "Geographic Information Sciences" or short "GISciences" (Bluemke

1 Most of the procedures in this book—geospatial data preparation, spatial linking, plotting of the maps, and analysis of the data—were realized using R. The corresponding code is available on request.

et al., 2017, 307), and also spatial methods are often called GIS methods or GIS techniques. Practice and training in GIS are still necessary to use geospatial data.

GIS makes it possible to process, analyze and visualize geospatial data. It allows, among others, accessing, manipulating, and converting of geospatial data and their corresponding metadata. Areas of application are manifold and range from the analysis of statistical relationships between two sets of geospatial data to the mapping and visualization of these relationships. By using GIS, researchers need knowledge about the procedures of spatial methods that this book exemplifies in different applications.

The main characteristic of all these methods is that they operate with locations. GIS identifies data points or observations through locations and conducts analyses based on locations; all methods rely on identifiers of locations, such as geo-coordinates. They do not only help to identify observations in the geospatial data but also to establish links between different sets of geospatial data. Chapter 2.4.1 gives more details on the specifics of the involved data in GIS. Overall, data that are targeted to be used in GIS must be georeferenced, a term and method that the following section introduces.

2.2 Georeferencing and Geocoding

Generally, georeferencing is the process of assigning standardized spatial references to data (Meyer & Bruderer Enzler, 2013, 323). These standardized spatial references are names or identifiers for spatial units, for example, administrative districts, municipalities or zip code areas (Hillmert et al., 2017, 270f). Assigning standardized spatial references to data depicts advantages for the processing and the analysis of data. Researchers, for example, can add information from auxiliary data sources to their data through matching spatial references; or they can statistically control dependencies between observations which cluster in space. Georeferencing provides a useful tool for a broad range of different purposes and applications.

At the same time, in the geosciences, such as ecology, the understanding of the term georeferencing is narrower: it depicts assigning *direct* spatial references to data, more specifically geo-coordinates (Rat für Sozial- und Wirtschaftsdaten — RatSWD, 2012, 11). In contrast to names and identifiers, geo-coordinates are coordinate points that define each location by a point on earth's surface. They are part of a Coordinate Reference System (CRS) which spans across the earth's surface and also respects its curvature (see Chapter 2.4.1). By assigning geo-coordinates to data, researchers can relate observations in space as they are part of a joined coordinate system, enabling explicit analyses based on this projection. The analyses include operations like the calculation of geodesic distances between points in the same coordinate system or

the calculation of areas around points of interest based on spatial proximity (Meyer & Bruderer Enzler, 2013, 27ff). Defining georeferencing as the assignment of geo-coordinates to data renders the use of spatial methods possible in the first place.

Thus, social science survey researchers who aim to use georeferencing for spatial analyses and spatial linking require geo-coordinates for their data. However, as most of the survey data are data with *indirect* spatial references, e.g., addresses of survey respondents, they first have to convert them into geo-coordinates. This operation of converting addresses into geo-coordinates is known as the method of geocoding that does not necessarily have to be implemented by the researchers. Instead, they can rely on automated services from providers such as Google, Bing, or, for the German case, the German Federal Agency for Cartography and Geodesy (BKG) which convert addresses into geo-coordinates with online tools (Zandbergen, 2014, 2). While converting addresses into geo-coordinates would also be possible manually, automated procedures are faster and less sensitive to error. Since the sampling of many surveys in social science research is address-based, geocoding is one of the most critical requirements for using georeferenced survey data.

2.3 Georeferenced Survey Data

Georeferenced survey data are survey data that contain direct spatial references in the form of geo-coordinates. Because geo-coordinates enable spatial methods, as described above, researchers can use geo-coordinates of respondents' locations to project them in a joined space and to relate them to each other. These geo-coordinates do not necessarily have to stem from housing addresses; they also can originate from geographic units, including administrative districts or municipalities. Previous work, however, suggested that the smaller the geographic scale of spatial units is, the better research questions toward the local context of people can be answered (Nonnenmacher, 2013). For this reason, this book investigates georeferenced survey data that contain geo-coordinates for particular small spatial units like neighborhoods or survey respondents' dwellings.

Meanwhile, georeferenced survey data are different from other georeferenced data. Data protection legislation in Germany forbids to store results from survey questions and personal information, such as addresses of respondents being, in one single file (Chapter 4.3 below discusses this issue in more detail). Because housing addresses make people identifiable, geo-coordinates from respondents' housing addresses depict personal information. In order to still work with these data, a common approach is to define a technical and organizational workflow that prevents confounding sur-

vey data and personal data (Schweers et al., 2016, 116). Data sensitivity is what makes georeferenced survey data different from other georeferenced data, and this is why researchers have to be cautious if they aim to spatially link these data to other data sources.

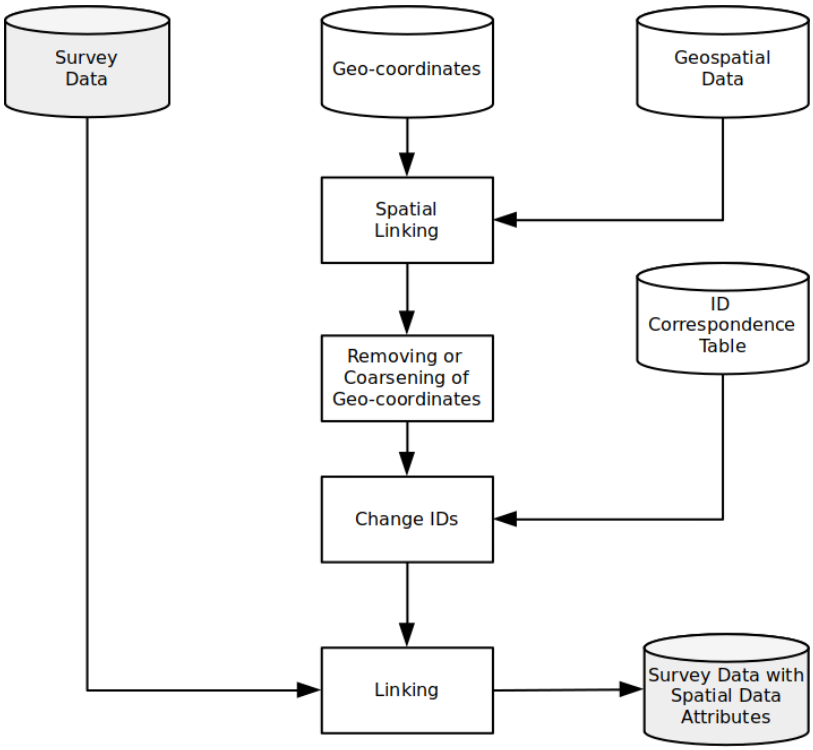


Figure 2.1: Organization of Spatial Linking Respecting Data Protection

Figure 2.1 shows an example of a workflow that allows spatial linking of georeferenced survey data (i.e., the enrichment of survey respondents' geocoordinates with geospatial information) by, at the same time, respecting data protection legislation. First, users of this workflow apply spatial linking procedures (see Chapter 2.5 and 5.2) to the survey respondents' geo-coordinates and some auxiliary geospatial data sources. In this step of the workflow, survey data and geo-coordinates are separated. The result is a new dataset that contains geo-coordinates and spatial data attributes from the auxiliary geospatial data sources. Second, the workflow subsequently envisages deleting the geo-coordinates from this new dataset, or coarsening the geo-coordinates by either permuting or aggregating them. The result of either deleting or coarsening is similar: it removes personal information from the data. Lastly, survey data and the created spatial information can now be linked via matching identifiers in a correspon-

dence table. This table is the sole way to establish a correspondence between survey data and personal data—research projects organize access to this table strictly. Applying all three steps of this workflow results in a dataset of survey data enriched with spatial attributes from auxiliary data sources.

To sum up, using georeferenced survey data is complicated due to legal barriers. Because these legal barriers involve even more associated issues, Chapter 4.3 discusses the background of data protection in more detail. Thus far, this section illustrates that georeferenced survey data require technical and organizational measures. However, using a well-organized workflow facilitates working with georeferenced survey data.

The legal barriers may be the reason for a still rather low prevalence of georeferenced survey data, at least in Germany. But this does not mean that producing georeferenced survey data in Germany is not possible at all. The following two sections introduce two survey datasets that were georeferenced based on survey respondents' addresses, the georeferenced GGSS 2014 and the GESIS Panel. The empirical applications in Chapter 6-8 display the analysis potential of these georeferenced survey data.

2.3.1 German General Social Survey 2014

ALLBUS

The German General Social Survey (GGSS) of the year 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015) is a two-stage disproportionate random sample of private households in Germany. All persons were at least 18 years of age by the time of the interview which they completed with a standardized questionnaire (CAPI – Computer Assisted Personal Interviewing) and two additional self-completion questionnaires (CASI – Computer Assisted Self-Interviewing). While the GGSS is conducted every two years aiming to monitor trends in attitudes, behavior, and societal change in the Federal Republic of Germany, it also implements changing core themes in specific waves. Table 2.1 gives an overview of some general sociodemographic characteristics of the sample.

The data of the GGSS 2014 were georeferenced in the project "Georeferencing of Survey Data" (GeorefUm) funded by the German Research Foundation at the GESIS Data Archive for the Social Sciences.[2] For this purpose, the addresses of the GGSS 2014 survey respondents were geocoded using the secure geocoding service of the Federal Agency for Cartography and Geodesy (BKG). This service complies with the German data protection legislation as it only processes address data on the fly, leaving no trace

2 https://www.gesis.org/en/research/external-funding-projects/archive/georefum/

of addresses on the BKG servers.

Table 2.1: Sociodemographics of the Georeferenced German General Social Survey 2014

	Mean / %	SD	Minimum	Maximum
Age	49.44	17.51	18	91
Gender (female)	49.24			
Income	1500.40	1560.78	0	60000
Education				
Low	13.20			
Medium	61.45			
High	24.40			
Number of Observations	3471			

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Together with the survey attributes the geocoded addresses of the GGSS 2014 comprise a georeferenced survey dataset.[3] In practice, however, managing these two sources is more complicated: as already described above, they are not allowed to be stored in one single dataset. The workflow in Figure 2.1, described earlier, displays a procedure that makes it possible to use the georeferenced GGSS 2014 for spatial linking purposes.

2.3.2 GESIS Panel

The GESIS Panel (GESIS - Leibniz Institute for the Social Sciences, 2017) is a probability-based mixed-mode access panel of 4,900 persons aged at least 18 years who complete their interviews every two months. What is unique about the GESIS Panel is that researchers can submit questions for the survey (Bosnjak et al., 2018, 104). For this reason, the researchers can draw on a large variety of different topics that the respondents are asked about. These topics range from political issues, health, and life satisfaction to environmental assessments and general as well as detailed sociodemographic characteristics. Exactly like the GGSS data, the addresses of the GESIS Panel used for spatial linking in this book were georeferenced in the GeorefUm project at GESIS. For the empirical application in Chapter 8 only a subset of the 2014 round of

3 By merging the GGSS 2014 survey data with the Sensitive Regional Data of the GGSS (GESIS - Leibniz Institute for the Social Sciences, 2018) this data can be replicated. Details about these data can be found in Klinger (2018).

the GESIS Panel is used. Some of the samples' sociodemographic characteristics are presented in Table 2.2.

Table 2.2: Sociodemographics of the Georeferenced GESIS Panel 2014

	Mean / %	SD	Minimum	Maximum
Age	46.15	14.08	18	73
Gender (female)	52.47			
Income (Categories)	11.69	2.84	1	17
Education				
Low	21.32			
Medium	35.12			
High	43.57			
Number of Observations	3882			

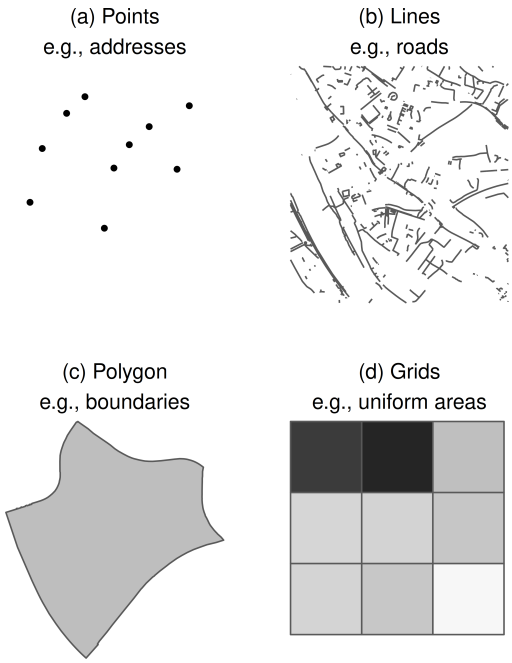
Data Source: Georeferenced GESIS Panel (GESIS - Leibniz Institute for the Social Sciences, 2017)

A comparison of the sample of the GGSS and the GESIS Panel shows some differences between their basic sociodemographics. Respondents of the GESIS Panel are younger and the sample consists of a higher rate of female respondents. The most striking difference concerns respondents' education: in the GGSS, the largest group are people with medium education. Instead, the GESIS Panel sample consists of a high share of people with low and high education. In later analyses with these data, it is crucial to adjust the estimates for these sociodemographic variables.

2.4 Geospatial Data

Generally, geospatial data hold information on geometries projected on earth's surface. This information either consists of the geometric extent, shape, and structure of the geometries or it also comprises additional contents associated with each geometry. Geospatial data are multi-dimensional and diverse, depending on the structure of the geometries as well as on the actual content of the corresponding attributes.

Accordingly, geospatial data can look rather different. Figure 2.2 shows a collection of geometries which are represented by specialized geospatial data formats (Meyer & Bruderer Enzler, 2013, 326). These include simple point or line geometries but also more complex polygon as well as evenly distributed grid structures. Their different shapes and structures correspond to different geographic counterparts in reality (Bivand, Pebesma, & Gómez-Rubio, 2008, 8ff).



Data Sources: OpenStreetMap / GEOFABRIK (2018) and City of Cologne (2014)

Figure 2.2: Different Geometries of Geospatial Data

These geographic representations are manifold. Points often depict fixed locations, such as hospitals, kindergartens, stores, or geo-coordinates of survey respondents' dwellings. Lines may be street or river courses, or airplane approach paths. Polygons, as mentioned, may comprise more complex structures, including neighborhoods, city districts or other administrative boundaries information, but also may result from measurements on air pollution or traffic noise. Lastly, rasters can contain all this information represented in evenly shaped grid cells.

Moreover, geospatial data are not only multidimensional and contain an extensive range of information. As each geometry within one and also between more data files can be presented in a common Coordinate Reference System (CRS) (see 2.4.1), it is possible to localize subsets of the data files with geo-coordinates and to put them in spatial relation to each other. It is this projection of the geometries in one coordinate system that enables these flexible operations.

To perform such operations researchers work with different data types and formats, deal with different geo-coordinates' projection logics, and access data through different channels. The following section gives more details on these areas which are the cornerstone of any work with geospatial data. Hence, the specifics of geospatial data

simultaneously are the basis for the method of spatial linking that section 2.5 presents.

2.4.1 Specifics of Geospatial Data

Data Types and Data Formats

Two classes of data types commonly embody the geometries of geospatial data: vector data and raster data (Meyer & Bruderer Enzler, 2013, 326). The associated file formats differ in the way they store the information, which in turn also impacts the processing of the data. The file formats' characteristics affect accessing and storing as well as visualizing and analyzing. Therefore, it is vital to elaborate on the differences between these geospatial data types.

Vector data hold information on points, lines or polygons (Sutton, Dassau, & Sutton, 2009, 10). In a point vector data file, each point represents a feature: an observation in the dataset which is described by a geo-coordinate. The same applies for line and polygon features, only that they contain more than one geo-coordinate. Regardless of what kind of geometry they represent, each of the features can hold more attributes in a separate data table (Sutton et al., 2009, 21ff). Points may contain information about the kind of dwelling a survey respondent lives in; lines may contain information on whether a street is a main road or byroad; polygons may contain information on immigrant or unemployment rates in a city district. Vector data are complicated because their features comprise information on their geometry and their attributes.

Raster data hold information on evenly shaped grid cells (Sutton et al., 2009, 47). In a raster grid file, at first, each cell of a data table represents an observation in the dataset (Meyer & Bruderer Enzler, 2013, 326). The cells are organized in their geographic order—from west to east and from north to south. Metadata such as the geographic extent of the data table, the size of each grid cell or the information about the CRS (see next section) turn this data table to a geospatial dataset (Sutton et al., 2009, 49ff). Without this metadata, a raster data file is not more than an image file that lacks georeferencing. The actual content of the cells, however, can contain all information vector data also represent. Raster data have a flat structure because grid cells contain information on their attributes, whereas information on their geography are stored globally as metadata.

The different structures of vector and raster data affect the purposes for which they are useful, ideally displayed by comparing the two data types one-to-one. First, vector data are expedient in cases in which geometries are complex or of a particular interest (Meyer & Bruderer Enzler, 2013, 326). While some technical routines can convert vector data to raster data (and vice versa) (Sutton et al., 2009, 53), this can also coarsen

the information. Instead, raster data types are useful in cases in which evenly shaped grids are appropriate to describe information in space. Second, in cases in which at least one observation in the data contains information about more than one attribute, vector data are more convenient to use as they separate information on geometries and attributes. In contrast, raster data generally hold information on one single attribute.[4] Third, as vector data are complex regarding their geometries, analyzing them is more complicated in comparison to raster data. In vector data, relating observations to each other typically involves the consideration of their geometry, where as in raster data, only the knowledge of their position in the data matrix is required.

Deciding between vector and raster data or even using both is a matter of the methods that researchers aim to apply. Chapter 5.2 describes different methods of spatial linking that require vector or raster data. Also, using either vector or raster data is not an exclusive effort: some operations require both data types in one analysis. Thus, after describing common elements of both data types in the following, Chapter 2.4.2-2.4.2 presents examples for vector and raster data of the empirical applications.

Coordinate Reference Systems (CRS)

Thus far, this book defined geo-coordinates as X and Y coordinate pairs that describe locations on earth's surface while also respecting its curvature. Earth's curvature as 3-dimensional surface, however, poses mathematical challenges when projecting it into a 2-dimensional space of X and Y coordinates (Sajevaan, 2008). Projection distorts geometries and relationships between geometries; projection is bargaining between what is useful for a specific purpose and what is the most accurate model of earth's surface in a specific region. For these reasons, numerous projection models exist that historically and practically result in different geometric extents and different geometric relationships.[5]

X and Y geo-coordinates that follow a specific projection system are part of a Coordinate Reference System (CRS) (Bivand et al., 2008, 84f). Definitions of CRS comprise

4 GIS can also define raster stacks or bricks (Hijmans, 2017, 2). They summarize different raster datasets that are geographically identical but hold different attributes. Raster stacks store the geographic information of multiple raster datasets only once and are more efficient to process than single raster datasets.

5 The website https://thetruesize.com demonstrates interactively and playfully, for example, how the commonly used Mercator projection distorts geometries of non-Western countries and continents. The Mercator projection is useful for navigation purposes because it provides the ability to draw straight lines which can be followed by vehicles on earth. However, its use for visually presenting the earth produces wrong expectations regarding the geometries of some regions.

line strings that contain information, among others, about the coordinate system's origin, the system's units (e.g., meters or feet), or its spheroid. In Listing 2.1, for example, the definition of the "ETRS89 / LAEA Europe" CRS is shown which more and more European providers of geospatial data use.

Listing 2.1: Example of an EPSG:3035 Definition File (Line Breaks and Indentions Are Included for Illustration Purposes)

```
PROJCS["ETRS89_LAEA_Europe",
  GEOGCS["GCS_ETRS_1989",
    DATUM["D_ETRS_1989",
      SPHEROID["GRS_1980",6378137,298.257222101]
    ],
    PRIMEM["Greenwich",0],
    UNIT["Degree",0.0174532925199433295]
  ],
  PROJECTION["Lambert_Azimuthal_Equal_Area"],
  PARAMETER["latitude_of_origin",52],
  PARAMETER["central_meridian",10],
  PARAMETER["false_easting",4321000],
  PARAMETER["false_northing",3210000],
  UNIT["Meter",1]
]
```

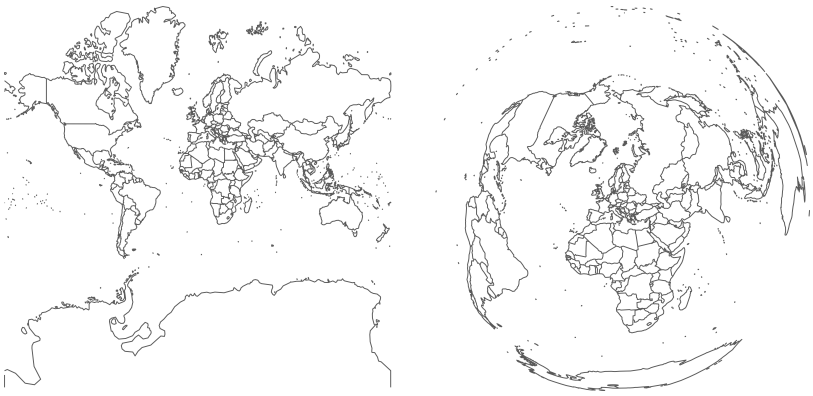
Any geospatial dataset requires such a definition of an CRS. As these definitions can be rather complicated, software for geospatial data provides tools to define CRS by specifying CRS codes, such as EPSG codes (Bivand et al., 2008, 85f).[6] The EPSG code for the "ETRS89 / LAEA Europe" CRS, for example, is EPSG:3035.[7] Thus, while CRS contains complex information on earth's curvature, applying them to geospatial datasets is straightforward.

The effect of different CRSs in the representation of geo-coordinates on the earth's surface is tremendous. Figure 2.3 shows the differences of creating a world map with the (Pseudo-)Mercator projection (EPSG:3857) that, for example, OpenStreetMap or Google use for mapping purposes and the ETRS89 / LAEA Europe Projection (EPSG:3035) that is prevalent in the European spatial data landscape. While a fictional drawn equator line in EPSG:3857 depicts a straight line, in EPSG:3035 it builds a curve resembling the curvature of the globe. Also, regions in the southern and northern regions of the earth are enormous in EPSG:3857, in the European-centric EPSG:3035

6 EPSG stands for European Petroleum Survey Group Geodesy, a project in which EPSG codes were invented: http://www.epsg.org/

7 http://spatialreference.org/ref/epsg/etrs89-etrs-laea/

projection they are smaller. None of the projections is wrong—they serve different purposes, but users of different CRSs have to keep in mind that with regards to the different mapping purposes results may change.



Data Source: Statistical Office of the European Union Eurostat (2018)

Figure 2.3: A World Map in (Pseudo)-Mercator Projection (EPSG:3857) on the Left and ETRS89 / LAEA Europe Projection (EPSG:3035) on the Right

The discussion on different projection models is extensive (Sheppard, 2005), and repeating all arguments in favor of a specific CRS is not subject of this book. Still, to work with geospatial data researchers need information on CRS. Furthermore, to combine different geospatial datasets their CRS have to match; otherwise their combination would result in wrong spatial relationships between geometries. While tools to automatically transform X and Y of a specific CRS into another CRS exist (Bivand et al., 2008, 88f), specifically for spatial linking efforts (see Chapter 2.5 and 5.2) matching CRS are vital.

Data Access via Web Services

Geospatial data can be enormous in file size. Analyzing the data is challenging and requires adequate computing power if users aim to handle all data at once. For this reason, in the geosciences, it is generally accepted to process the data based on database accesses that request chunks of data for a specific purpose. Working with geospatial data on a request basis decreases computing time and increases efficiency.

Of those data providers which decided to offer their data on the internet, some even offer programming interfaces (Meyer & Bruderer Enzler, 2013, 325). Users of the data do not have to download their data as a data dump in a specific geospatial data format. Instead, they can send requests for individual geo-coordinates or bounding boxes—

rectangular geographic areas defined by minimum as well as maximum longitude and latitude values—and receive the underlying data (Percivall, 2016). Not all of these web services offer data to download as some provide interfaces solely for visualization purposes. If data privacy concerns allow processing georeferenced survey data with web services (see Chapter 4.3), the cases in which these services offer data to download present an alternative to processing geospatial data all at once.

While in Germany more and more geospatial data are offered via web services[8], of all the data required for this book, only the data of the Leibniz Institute for Urban and Regional Development could be obtained via such a service. A Web Coverage Service (WCS) offers a programming interface to their raster data that enables users to request data for specific bounding boxes.[9] All other data either had to be acquired individually (Schweers et al., 2016, 109ff), or exist as a data dump on the internet (Müller*, Schweers, & Siegers, 2017, 13ff). In any case, because European initiatives expand the landscape of administrative geospatial data (Schweers et al., 2016, 108), researchers can expect that the supply of geospatial data through web services will increase in the next years.

The following sections introduce the different geospatial datasets for the empirical applications in Chapters 6-8. They stem from a diverse set of disciplines ranging from acoustic engineering to social statistics and the ecological development sciences. Each dataset requires training in dealing with the data collection and the measurements of each discipline. As the general remarks about geospatial data above may suggest, the geospatial data for this book represent a small extract of what is conceivable for social science research in even more research applications. The applications in this book are inevitably research examples and do not represent an exhaustive analysis of all possible applications.

8 See, for example, https://www.geoportal.de/EN/

9 The spatial linking of the downloaded data yet is still up to the researchers. As this demands knowledge in software and data, a project funded by the German Research foundation called "Social Geospatial Research Data Infrastructure" (SoRa) builds a tool for social science researchers to conduct spatial linking procedures of these data with georeferenced survey data: www.sora-projekt.de

2.4.2 Geospatial Data Sources

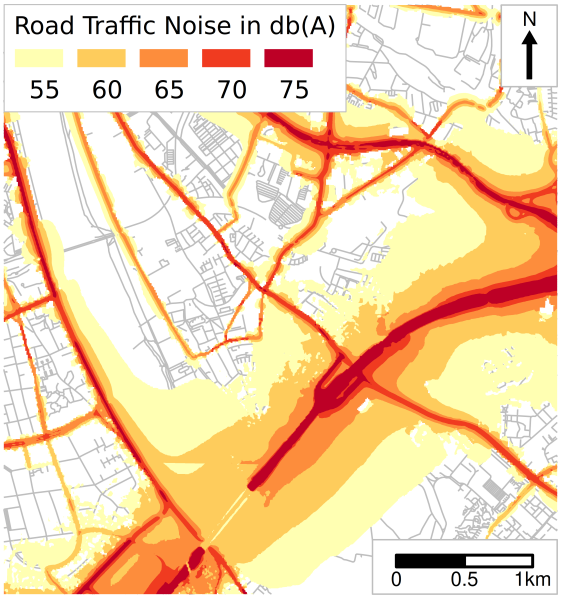
Road Traffic Noise

Some data which are attractive for research are geospatial data that the European Member States collected in correspondence with the *Environmental Noise Directive* (2002/49/EC) of the European Union (EU) (European Parliament & European Council, 2002). One of these collections deals with road traffic noise of the year 2012 and provides the basis for the empirical application in Chapter 6 (German Environmental Agency / EIONET Central Data Repository, 2016). Generally, the directive obligates members of the EU to collect data on noise sources originating from industries and air, rail traffic as well as road traffic. In this directive, noise is defined as an exceeding level of sound pressure measured in A-weighted decibels (dB(A)), either 55 dB(A) by day or 50 dB(A) by night. The data for this book capture road traffic noise as day-evening-night-mean levels ($\geq$ 55 dB(A)).

The directive limited the collection of data to road traffic noise sources that were frequented by a specific amount of vehicles: the data only cover main roads which carry more than three million cars a year, that is five cars per minute. For this reason, plotting the data on a map results in a large number of non-covered areas for which no road traffic noise measurement is available. Figure 2.4 displays an example of the road traffic noise data for a map section of an area in the city of Cologne. All over the map, the white areas show the effects of the missing data collection—in these areas, no measurements of road traffic noise are available.

Concerning the issue of non-covered areas, one may argue that these areas are not noisy by definition of the EU directive. While there may be loud and heavy traffic on smaller roads as well, road traffic noise is not defined by temporary but by enduring levels of dB(A). Studies on environmental noise and its effect on people, e.g., show that mainly noise at night time affects people's health (Bocquier et al., 2014; Schmidt et al., 2013), which is also represented in the day-evening-night-mean. Thus, the road traffic noise data with its long-time measurements accurately represent traffic noise and its associated consequences, albeit being censored and not covering large areas.

As the empirical application in Chapter 6 shows, this particular structure of the road traffic noise data has consequences for conducting research. The data consist of spatial units or geometries of non-uniformly shaped polygons with missing measurements for byroads. Population survey data, on the other hand, often consist of a random sample of the population—people who live in main roads and in byroads. As a consequence, combining these data results in a considerable number of missing



Data Sources: German Environmental Agency / EIONET Central Data Repository (2016) and OpenStreetMap / GEOFABRIK (2018)

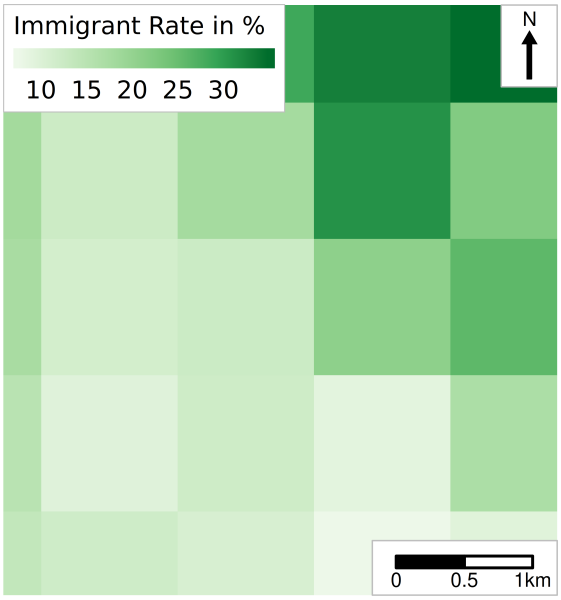
Figure 2.4: Road Traffic Noise Data in an Area of the City of Cologne

road traffic noise measurements for the sample since a large number of people live in areas with low road traffic noise exposure. Studying the effects of road traffic noise in a population survey leads to statistical issues because of small sample sizes.

German Census 2011

One of the most comprehensive and also publicly available data sources for the whole extent of Germany is the German Census 2011 which is used in the empirical application in Chapter 7. The 2011 European Census Regulation (European Parliament & European Council, 2008) obliged members of the EU to collect these data to provide information on the sociodemographic composition of the EU population. Moreover, the German Federal Agency of Statistic (destatis) gathered the German Census 2011 data on a rather small scale. For example, the data contain information on immigrant rates used by the empirical application for uniformly shaped 1 km^2 grid cells for the whole extent of Germany—information that authorities normally only provide on the municipality level. While today the German Census 2011 data have aged, they still depict exceptional and compelling data for research in Germany.

Figure 2.5 displays the structure of these data for a map section of an area in the city of Cologne at the example of immigrant rates (Statistical Offices of the Federation and the Länder, 2016). In comparison to the map of road traffic noise before, this map covers the whole area of the section—no areas are left out (non-rectangular grids are artifacts which results from cutting out the map from a larger dataset). As Chapter 5 describes in more detail, uniformly shaped grid cells which cover the whole area have advantages for spatial linking and analyzing these data.



Data Source: Statistical Offices of the Federation and the Länder (2016)

Figure 2.5: Amount of Immigrants in 1 km^2 Grid Cells in an Area of the City of Cologne

Other sociodemographic attributes are available comprising information on different age groups, household sizes, vacancy rates, or flat sizes. Lately, destatis published an even more extensive list on a smaller geographic scale of 100 meters $\times$ 100 meters. The data now include, for instance, information on building types, families and new sociodemographics such as religious denomination. Due to data permutation methods necessary for data protection, these data, however, must be treated with caution (Statistische Ämter des Bundes und der Länder, 2011, 67ff). Therefore, the empirical applications in this book do not use them as a data source for the analysis.

Structurally, the German Census 2011 data are raster data. Raster data have the advantage of being smaller in file size compared to other geospatial data formats and, at the same time, are computationally efficient. For these reasons, even though the German Census 2011 data consist of 557,256 grid cells, the file size of these data is on

average just ~12 Megabytes for each file. Lastly, as the data extend the whole area of Germany, there are no issues of non-matching cases as with the road traffic noise data because of randomly sampled populations in survey research.

Soil Sealing

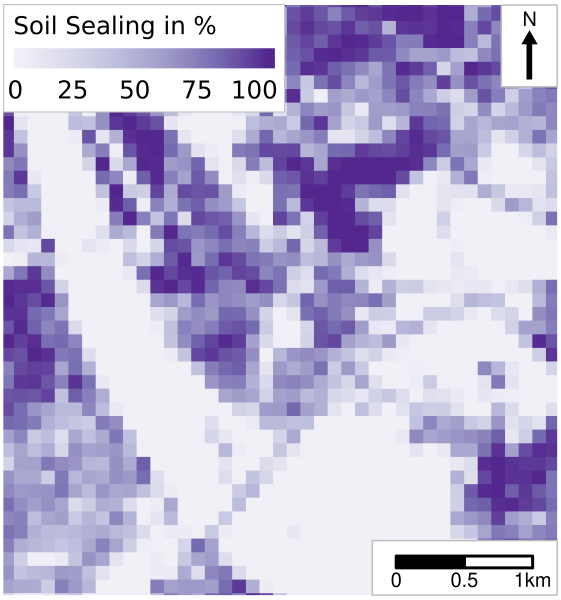
Another comprehensive set of geospatial data for the whole extent of Germany is land use data from the Leibniz Institute for Urban and Ecological Development (IOER) which are used in the empirical application of Chapter 8. Their Monitor of Settlement and Open Space Development (IOER Monitor) provides access to a broad range of different land use indicators ranging from settlement structures to land quality and relief information.[10] This book's empirical applications use one of these indicators: soil sealing of the year 2014 (Leibniz Institute of Ecological Urban and Regional Development, 2018).

Soil sealing is the air and water tight coverage of an area by buildings and traffic areas. As a result of this coverage, water cannot seep away, and it affects and disturbs the gas interchange with the atmosphere. These environmental impacts are lasting: removing soil sealing is costly and in some cases even hard to realize.[11] Soil sealing, an admittedly abstract land use indicator, depicts an environmental hazard in areas with a high density of buildings and traffic areas.

Figure 2.6 displays the structure of soil sealing data in another map section of the city of Cologne. The data mainly cover areas with high densities of big roads and buildings, which depicts a more nuanced distribution than that of the 1 km^2 grid cells of the Census data. The reason is that the data are far more detailed as they comprise 100 meters $\times$ 100 meters grid cells. Already a small map section of soil sealing data contains a broad set of information.

As soil sealing has increased in Germany in the last decades, it affects people and society as a whole. Between the year 1992 and 2011, for example, the amount of soil sealing has grown from 5.3 % to 6.2 % (excluding data from Saxony-Anhalt).[12] While in recent years the annual rate was less distinct, land use through soil sealing remains a severe political issue. Areas that deal with large amounts of soil sealing miss, for example, other recreational and free areas such as green spaces. These areas were shown to have serious effects on people's well-being (Gidlöf-Gunnarsson & Öhrström, 2007),

10 http://www.ioer-monitor.de/
11 https://www.umweltbundesamt.de/en/topics/soil-agriculture/land-a-precious-resource/paving-construction
12 https://www.umweltbundesamt.de/sites/default/files/medien/384/bilder/dateien/2_abb_anteil-suv-gesamtflaeche-d_2013-10-02.png



Data Source: Leibniz Institute of Ecological Urban and Regional Development (2018)

Figure 2.6: Water and Air Tight Coverage of Land in 100 Meters × 100 Meters Grid Cells in an Area of the City of Cologne

stress processing (Thompson et al., 2012), and health (Guite, Clark, & Ackrill, 2006). Again, for these reasons soil sealing represents an environmental hazard worthwhile to study in social science applications.

Structurally, the land use data from the IOER are also raster data that cover the whole extent of Germany. Also, because the IOER used no data permutation methods, all 100m × 100m grid cells represent valid and reliable cases that can be used for spatial linking purposes. In total, the data contain 67,108,864 grid cells.

Administrative Boundaries

In the applications as well as in the illustrations of this book, several other geospatial data sources are used, most of them being administrative boundaries of different types and different area sizes. The following briefly introduces these data sources.

In the case of city districts, the data stem from the open data portal of the city of Cologne (City of Cologne, 2014). This portal offers a large amount of data on different topics from different city authorities—geospatial data represent only one of them. In particular, the data used in this book provide detailed insights into the shapes and the extent of city boundaries. They are useful to compare the geographic scales of

different geospatial datasets.

Maps with geospatial data from the OpenStreetMap (OSM) project visualize roads and buildings (OpenStreetMap / GEOFABRIK, 2018). The website of GEOFABRIK offers a large amount of already prepared geospatial data for geometries retrieved from OSM, e.g., in the ESRI shapefile format (ESRI, 1998). Using GEOFABRIK represents a convenient way of accessing the data from OSM. Like the data from the open data portal of the city of Cologne, OSM data in this book are used to compare the geographic scales of the other geospatial datasets.

One last source of geospatial boundaries data are data on municipality boundaries. They stem from the Federal Agency of Cartography and Geodesy and are also available as a free download (Federal Agency for Cartography and Geodesy, 2018). Not only are these data used for visualization purposes but also to locate survey respondents within single municipalities. The localization is useful to control spatial dependencies between respondents that cluster in municipalities (see Chapter 5.3). In all empirical applications in this book, municipalities data are used as part of the statistical models.

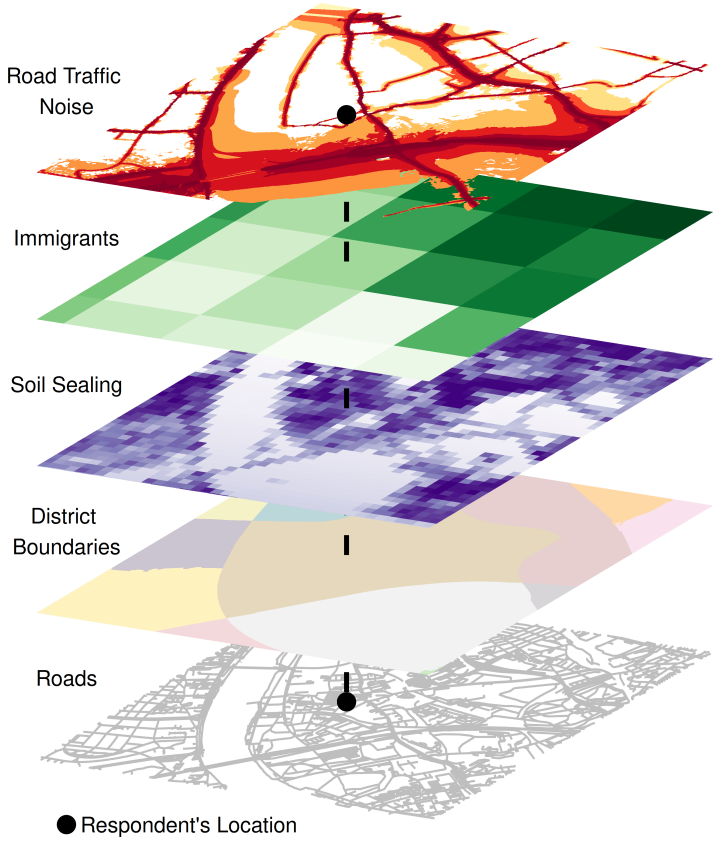
2.5 Spatial Linking

Apart from the general discussion on georeferenced data, spatial linking methods are central in this book. Spatial linking methods depict techniques of combining two georeferenced data sets, also called spatial overlay or spatial join (Bivand et al., 2008, 140ff). They are crucial to the use of georeferenced survey data if researchers aim to combine them with data from other scientific disciplines. For this reason, Chapter 5.2 presents spatial linking methods in all detail. This section introduces the basic idea of spatial linking.

Spatial linking is a specific technique of data linking. Instead of using standardized identifiers in the form of names or distinct numbers, the linking happens by projecting georeferenced data into a joined coordinate space. In the simplest case, users of spatial linking methods retrieve information based on identical locations, e.g., between survey respondents' geo-coordinates and other spatial information, such as road traffic noise at their dwelling. Hence, spatial linking involves the definition of one focal dataset to which information from another dataset is added.

In the example of linking road traffic noise to survey respondents' geo-coordinates, the focal dataset is the respondents' geo-coordinates. One research purpose of this effort may be to see whether road traffic noise affects people's health in comparison to people who live in quiet spaces (Shepherd, Welch, Dirks, & McBride, 2013). Spatial linking uses the locations of the respondents, extracts the road noise level value at this

location and links this information to the geo-coordinates. In combination with the actual survey data, researchers then also can analyze whether road traffic noise predominately affects people of specific social groups, such as low-income groups. Thus, spatial linking with survey data uses combined information from geospatial information and the answers of people in standardized surveys.



Data Sources: OpenStreetMap / GEOFABRIK (2018), City of Cologne (2014), Leibniz Institute of Ecological Urban and Regional Development (2018), Statistical Offices of the Federation and the Länder (2016), and German Environmental Agency / EIONET Central Data Repository (2016)

Figure 2.7: Spatial Linking as Projection of Multiple Layers into a Joined Coordinate Space

This method of spatial linking is just one use case, whose results involve binary outcomes—linking successful or linking not successful—and the corresponding values, such as road traffic noise values. The advantage of a projection in space is that spatial linking can establish relationships between two data sources based on their spatial relation to each other. As already presented in Chapter 2.4.2, the road traffic noise data depict censored data, i.e., mapping the data results in large areas of non-

measured spaces. Spatial linking to population survey data leads to notable numbers of missing values. Alternative methods of spatial linking, for example, the calculation of geographic distances to noise sources, may present an option to navigate such issues: researchers can use them as a proxy measure of noise exposure. The results of such methods of spatial linking are no longer binary and show the general flexibility of these methods.

Figure 2.7 displays the variety of spatial linking. The projection into a joined coordinate space combines data from different sources and of different structures: road traffic noise as polygons, immigrant rates as 1 km^2 grid cells, soil sealing as 100 meters $\times$ 100 meters grid cells, and boundaries of city districts as polygons. The location marker on each of these layers represents a direct locational correspondence to the marker on the bottom layer: the location of a fictional survey respondent's address, illustrated by a map of roads. This direct correspondence either enables a one-to-one spatial linking to the focal data of the survey respondent's location or the inclusion of surrounding spatial information via geographic distances and other measures (described in Chapter 5.2).

Table 2.3: Spatial Linking Combinations Between Survey Data and Geospatial Data Across the Empirical Application (I-III)

	GGSS	GESIS Panel	Method
Road Traffic Noise	I		By Location (5.2.1), Geodesic Distances (5.2.4)
Immigrant Rates	II		Focal Linking (5.2.3)
Soil Sealing		III	Buffer (5.2.2)
Administrative Boundaries	I, II	III	By Location (5.2.1)

This book's empirical Applications I-III in Chapter 6-8 comprise different combinations of the georeferenced survey data and the geospatial data that were introduced in the previous two sections (2.3 and 2.4). Table 2.3 displays them by also indicating the method of spatial linking that was applied to create the combinations. Each empirical application uses one of the geospatial data sources: road traffic noise, immigrant rates, and soil sealing, whereas the administrative boundaries are used in all of them. Although it is possible to combine information from more than three data sources at a time via spatial linking, the empirical applications in this book keep it straightforward.

2.6 Translating Space into Socially Relevant Context

Working with georeferenced survey data is the practice of a spatially integrated social science (Goodchild et al., 2000). This effort involves new data types or other statistical models (Hillmert et al., 2017) as well as considerations of theoretical concepts. Albeit working with georeferenced survey data involves various methodological efforts, the implications of georeferenced survey data for theory are crucial. By using georeferenced survey data and spatial linking methods, what kind of information confronts researchers in their research?

Thus far, previous research thoroughly discussed the theoretical implications of a spatially integrated social science. Some of the arguments draw on the constructivist question whether spaces are socially constructed or something that exists independently from human judgment (Baur, Hering, Raschke, & Thierbach, 2014). By combining Critical Theory and geography, moreover, some researchers even ask whether the construction of maps displays and passes societal inequalities and power structures—a strand of research known as Critical GIS (Sheppard, 2005). Regardless of the stance of this work in this discussion, it is uncontroversial that constructing measurements for neighborhood influences indeed assumes some theoretical models of people's neighborhood perception. Comparable to survey scales, measures from spatial data should reflect the survey respondents' perception of their spatial surrounding (Bluemke et al., 2017, 311).

Meanwhile, space is a rather general term. It subsumes all actions and states that happen or exist at a specific location during a specific time (Baur et al., 2014, 15f). More specifically in this regard, locations can be small neighborhoods consisting of a few houses surrounded by large roads, city districts, the area of a municipality, or even whole states. Another critical component of this definition is time: space changes over time, maybe to its extent, its constitution or even its meaning for people. Space, defined as relevant locations for people that influence their life, may not be stable for all time.

On a practical note, for the work with georeferenced survey data, it is ideal if the information from spatial data match with the respondents' location during the time of the interview. Otherwise, researchers would link information displaying different time points and, thus, draw wrong inferences. Information about spaces would become meaningless and would be left as being only places (Baur et al., 2014, 11). Potential relationships that researchers find in these data were spurious. Because of these reasons, it is necessary to find matches between space and time to operationalize people's perception of their spatial surrounding.

Going back to the conceptual framework of space, the definition of space yet is still

rather general. Spaces are locations that are relevant in a specific way, for people, administrations, or the whole society. They define specific geographies either by borders and areal extents (e.g., neighborhoods, or municipalities), or landmarks (e.g., rivers and mountains, but also points of interest such as shopping centers, kindergartens, and train stations). Thus, the actual answer to the question why they are relevant for people varies with their content: municipalities affect people because of policies, infrastructures and the people who live there; landmarks can limit or provide access to resources or other locations within an area. Relevant spaces, therefore, depend on the geography as well as the content they provide.

By presenting and discussing the work of others, Chapter 3 exemplifies that this general definition of space makes sense for social science studies. Research in which the geographic scale of spatial information varies, for example, helps to find out to which extent neighborhoods are still relevant for people (Sluiter et al., 2015; Tolsma & van der Meer, 2017). Another line of research that analyzes the role of boundaries within spaces emphasizes the fine-grained nuances in space and its influence on people (Legewie & Schaeffer, 2016). In general, the choice of how researchers operationalize and measure space depends on the actual research question.

To narrow down this broad scope of a spatially integrated social science, this work makes one necessary constraint concerning the examples and its applications: it will concentrate on research with a geographic scale below or equal to the level of municipalities. The literature on spatially integrated social sciences is already extensive, especially regarding research questions and applications that focus on larger spatial units such as administrative districts, or states. Previous work widely discussed the implications for statistical and theoretical inference stemming from such applications (Nonnenmacher, 2013; Schmidt-Catran & Fairbrother, 2016). Furthermore, the demand of the social sciences for using spatial information on a particular small spatial scale increased (Schweers et al., 2016), and is actively advertised (Bluemke et al., 2017; Nonnenmacher & Friedrichs, 2011). As the demand for small-scale georeferenced data and its methodological implications lack a thorough discussion, this work will concentrate on specific considerations regarding these data.

3 Applications for Georeferenced Survey Data

After presenting the basic terms of and general ideas for georeferenced survey data, the most important questions remain: Why should social science researchers use georeferenced data in the first place? What are the research applications that justify bothering with the interdisciplinary effort of a spatially integrated social science? Moreover, what has past research shown, what are the remaining open questions? Besides this book's general contribution to these questions, the present chapter gives some first answers by highlighting studies of others. It discusses theoretical implications and reoccurring methodological themes.

3.1 Exemplary Research Fields

Researchers use small-scale spatial data in a broad range of different areas of the social sciences. To name just a few, they range from health (Bocquier et al., 2013; Oiamo et al., 2015; Saib et al., 2014), social inequalities (Downey, Crowder, & Kemp, 2016), political behavior and migration (Dill & Jirjahn, 2014; Förster, 2018), values and attitudes (Klinger et al., 2017) to educational research (Ainsworth, 2002; Crowder et al., 2011; Weßling, 2016). The following presents some of these studies to exemplify their contribution to the field. Because of their mere number, the examples are limited to the fields that this book's empirical applications belong to: health, social inequalities, and political attitudes research. Moreover, the number of studies even in the fields presented here is big. The examples, therefore, do not depict a complete list of applications with georeferenced data. Instead, they represent illustrative examples that emphasize the utility but also the peculiarities of this research.[13]

3.1.1 Health and Neighborhoods

Space has been of interest in social science health research already for a long time. Ecological health models conceptualize people's health as influenced by individual and contextual factors, including social networks, neighborhoods, or even state policies (Berkman, Glass, Brissette, & Seeman, 2000; Pescosolido, 2006). These models assume that people themselves shape space by comprising neighborhoods with their

13 For systematic reviews, for example, on the topic of neighborhood effects, please refer to the admittedly aged but still relevant work of Dietz (2002) or Sampson et al. (2002).

actions or their composition of characteristics. Spaces, in turn, affect people through geographic proximity in the simplest case—people compose spaces, and spaces influence their health. Moreover, also daytime and moving between spaces, such as commuting to workplaces, are considered as an essential mechanism of how spaces affect people's health (Rainham, McDowell, Krewski, & Sawada, 2010). Space in health research with its corresponding theoretical models is a non-static criterion, leading to a broad range of different applications.

For example, over the last two decades, an extensive body of research on neighborhoods' effects on mental disorders has emerged (Cutrona, Wallace, & Wesner, 2006; Haines, Beggs, & Hurlbert, 2011; Ross, 2000). Researchers generally found that people who live in disadvantaged neighborhoods report higher mental distress (Mair, Diez Roux, & Galea, 2008); thus, disadvantages of neighborhoods become individual disadvantages. What explains these disadvantages were often social factors such as social support between neighbors (Kim & Ross, 2009; Kruger, Reischl, & Gee, 2007; Thorlindsson, Valdimarsdottir, & Hrafn Jonsson, 2012). Also, factors such as crime or violence are prominent themes in research (Curry, Latkin, & Davey-Rothwell, 2008; Latkin & Curry, 2003), which then again connects to the socio-economic structure of neighborhoods (Sampson et al., 2002). The example of mental health already shows how research leads to a variety of applications when researchers consider spatial attributes in their research question.

A few years ago, however, Ross and Mirowsky (2008) critically asked to what degree this link between neighborhoods' socio-economic factors and people's health comprises actual contextual effects. By using US census tracts as geographic units, they compared characteristics of these geographic units and individual characteristics with regards to their effect on people's health. They found a stronger association between the individual-level variables than between the census tract characteristics and people's health. Neighborhood effects were accordingly rather small in comparison, but, in line with other research (Lofors & Sundquist, 2007), they remained statistically significant. Ross and Mirowsky (2008) concluded that many findings in the literature are not caused by actual contextual effects but by the compositional effects of individual people in the neighborhood.

Because of this critical relation between neighborhood context and neighborhood composition, recent studies refined their methods for investigating neighborhood characteristics and health. Studying also the moving behavior of people into neighborhoods, Termorshuizen, Braam, and van Ameijden (2015), for example, reported an association between ethnic density and suicide risk among ethnic subgroups in the Netherlands. They found that higher rates of specific ethnic subgroups lower the risk

of committing suicide for members of these groups, mainly within non-Western minorities. The authors explained these findings with fewer experiences of perceived discrimination and more buffering through social support of people's own group members. While ethnic segregation varies between Western societies (Schönwälder & Söhn, 2009), findings from countries with more strict immigration policies, such as Canada, yielded similar results (Jurcik, Ahmed, Yakobov, Solopieieva-Jurcikova, & Ryder, 2013). These examples show that not all questions in the complex relationship between neighborhoods and people's health have been answered yet.

Moreover, rather than just concentrating on socio-economic contexts, other studies on neighborhoods and health also considered different neighborhood characteristics as well. Researchers discussed such environmental influences that intrude people's lives on an everyday basis, including environmental noise (Passchier-Vermeer & Passchier, 2000). Recent studies corroborated links to hearing problems (Basner et al., 2014), cardiovascular diseases (Babisch et al., 2014) and even diabetes (Sørensen et al., 2013). As some of the noise sources are prevalent among the general population, such as road traffic noise, environmental influences are also part of an ongoing societal debate.[14]

While the results regarding physical health and road traffic noise are noteworthy, the findings regarding mental illnesses such as depression or anxieties (Stansfeld, Haines, Burr, Berry, & Lercher, 2000) are ambivalent. The same holds true for subjective measures of health; many studies hardly find any effects at all. Roswall et al. (2015), for instance, who reported weak associations between road traffic noise and mental health, were able to explain their findings solely by lifestyle factors. Then again, Hardoy et al. (2005) already showed earlier that under consideration of subclinical symptoms, and even though their study was on air traffic noise, associations can persist. Accordingly, the ambivalence of findings leaves room for questioning possible explanations.

What a large set of explanations identify as a crucial factor for understanding the ambivalent findings is environmental noise perception (Jakovljevic, Paunovic, & Belojevic, 2009; Nitschke, Tucker, Simon, Hansen, & Pisaniello, 2014; Oiamo et al., 2015). Heritier et al. (2014) demonstrated that environmental noise perception mediates the effects of road traffic noise on health. Shepherd, Welch, Dirks, and Mathews (2010) found vanishing effects of road traffic noise after controlling additional sociodemo-

14 For example, municipalities such as the city of Cologne offer information on their websites that explicitly ask for the participation of their citizens to reduce noise: https://www.umweltbundesamt.de/en/topics/soil-agriculture/land-a-precious-resource/paving-construction

graphic variables, although it was mediated by noise annoyance and sleep disturbance. This twist in the study of environmental noise and health illustrates how the effects of spatial attributes also depend on their individual perception.

In spite of the empirical application in Chapter 6 also uses the example of road traffic noise, it just depicts one example of environmental stressors. Other environmental stressors that affect people's lives are manifold, ranging from air pollution (Marques & Lima, 2011; Rüttenauer, 2018), population density (Miles, Coutts, & Mohamadi, 2012) and land-use deterioration (You, Spoor, Ulimwengu, & Zhang, 2011) to missing access to recreational green areas (World Health Organization, 2016). Many of the findings directly relate to studies of residential and segregational patterns because exposure to environmental stressors is not necessarily a matter of voluntarily moving to, e.g., neighborhoods (Boes, Nüesch, & Stillman, 2013). In the next section, the perspective on environmental hazards, therefore, expands to social inequalities and their determining factors.

3.1.2 Social Inequalities and Environmental Hazards

Generally, disadvantages between neighborhoods can also result in social inequalities, meaning that neighborhoods affect people as individuals and also as a group (Braubach & Fairburn, 2010). This notion implies that people are not randomly sited in disadvantageous neighborhoods. They may also compose disadvantages through their sociodemographic characteristics (Ross & Mirowsky, 2008), as mentioned earlier, but segregational opportunities such as low housing prices could likewise influence migration into a specific neighborhood. Accordingly, social inequalities research emphasized the role of segregation of people in disadvantaged neighborhoods.

However, what defines this inequality in particular? As the examples from health research already have shown, neighborhood characteristics are often operationalized as aggregated measures of social indicators. Studies, for instance, may ask for the interplay between people's employment status and neighborhood unemployment rates, hypothesizing that if unemployment rates are high, people may be at risk to be unemployed as well. Although such research questions create theoretical challenges regarding contextual vs. compositional effects (Ross & Mirowsky, 2008) or issues of endogeneity (Dietz, 2002), in fact, they are a prominent theme in social science research (Curry et al., 2008; Haines et al., 2011).

Moreover, regardless of the associated issues, it is well known that people with specific characteristics, such as ethnicity A or B, cluster in neighborhoods (Schönwälder & Söhn, 2009) that have, in turn, their own environmental characteristics. There may

not be an interplay between individual and aggregated disadvantages, e.g., unemployment on the aggregated level does not produce unemployment on the individual level for one specific group. However, other environmental influences can cluster in these neighborhoods and produce inequalities (Braubach & Fairburn, 2010), e.g., regarding air pollution exposure (Crowder & Downey, 2010) or environmental noise (Bocquier et al., 2013). In combination with the health hazards which they impose (see the previous section 3.1.1), these environmental influences can accumulate disadvantages (Oiamo et al., 2015). This line of research shows that not only sociodemographic factors affect each other but also that other non-social factors may reveal interactions with people's lives.

For example, Rüttenauer (2018) analyzed if air pollution hazards of industry sites mainly affect neighborhoods with high rates of immigrants in Germany. By using aggregated German Census 2011 data (see Chapter 2.4.2) and location data of industry sites, he found that indeed industry sites locate themselves in neighborhoods with high shares of immigrants. This finding was even stronger among neighborhoods that are spatially clustered with neighborhoods which have similar immigrant rates. Moreover, Rüttenauer reported a difference between urban and rural areas in the sense that the effects are solely visible in urban areas. This association between clustering of neighborhood characteristics and environmental hazards motivated an extensive list of studies asking for the causal mechanisms on the micro level.

For the US societal context, Crowder and Downey (2010) studied whether socioeconomic factors can explain differences in exposure to air pollution between ethnic groups. They found that inequalities persist, even after controlling factors such as education or household income. Likewise for Germany, Best and Rüttenauer (2018) rejected the general hypothesis that income reduces environmental noise annoyance. In their study, using longitudinal data from the SOEP, income did not comprehensively decreased such annoyances, even after people were moving to new neighborhoods. Environmental inequalities persist, and social science researchers rightly ask for the underlying mechanisms of these findings.

Past research accordingly discussed environmental inequalities in regards to different theoretical considerations of residential segregation. Broadly, these considerations are either arguments stressing the socio-economic mechanisms of residential segregation, such as neighborhood (Zwickl, Ash, & Boyce, 2014) and individual income (Hoffmann, Robra, & Swart, 2003); or arguments that emphasize the social processes behind residential segregation, for example, whether more impoverished people move into polluted neighborhoods or whether pollution sources locate themselves in neighborhoods with more impoverished people (Best & Rüttenauer, 2018; Lersch,

2013). The empirical application in Chapter 8 compares these approaches with regards to land use hazards exposure and shows how georeferenced survey data can help to dissolve some of the theoretical challenges.

The next section expands this view on sociodemographic factors, such as ethnicity, to the overall perception of immigrants in society. A large number of theories emphasized the role of spatial relationships and neighborhoods, for example, in the study of prejudices against immigrants. The following, therefore, will exemplify the use of georeferenced survey data for this line of research.

3.1.3 Attitudes Towards Migration

Political attitudes and neighborhood conflicts research stands out with regards to its use of georeferenced data. By using small-scale spatial information, researchers can draw on detailed neighborhood characteristics. In comparison to typical applications that often operate with data on higher spatial levels and pose risks of ecological fallacies (Wu, 2007), small-scale geospatial information allows analyzing dynamics between and within neighborhoods. Particularly neighborhood conflict research can profit from such detailed data because of their rather demanding hypotheses.

Useful examples are applications that either use the *Intergroup Threat Theory* (Stephan, Oscar Ybarra, & Morrison, 2009) or the *Contact Theory* (Allport, 1954) to explain neighborhood conflicts. Both theories define specific hypotheses on a person-by-person level while also considering the neighborhood as a social entity. Following the Intergroup Threat Theory, people feel threatened by the presence of foreign people and fear a competition about socio-economic resources or cultural dominance. Instead, following the Contact Theory, non-superficial contact diminishes such feelings of threat. Although recent studies and the empirical application in Chapter 7 of this book aim to combine both theories (Klinger et al., 2017; Martig & Bernauer, 2016), they require data on a particular small geographic scale.

Effectively, recent research has been flourishing because of small-scale spatial data that became available. In their study of ethnic diversity in the neighborhood and its link to social trust, Dinesen and Sønderskov (2015) used neighborhood-level data within a starting radius of 80 meters around survey respondents' dwellings. At first, they found a positive effect between ethnic fragmentation and social trust. However, complementary to other research in this area (Sluiter et al., 2015), this effect vanished more the larger they varied the size of the radii around the respondents' locations—the measurement of the neighborhoods' influence got less precise. In general, social isolation as cause of ethnic heterogeneity is discussed as an explanation for these find-

ings (Dirksmeier, 2014, 843). This social isolation, in turn, decreases social capital as people have less and less supportive interaction with each other (Förster, 2018; Putnam, 2007).

The *Contested Boundaries Hypothesis*, introduced by Legewie and Schaeffer (2016), is an alternative explanation for poor social trust and resulting neighborhood conflicts. It states that neighborhood conflict "arises at poorly defined boundaries that separate ethnic and racial groups" (Legewie & Schaeffer, 2016, 126). In contrast to existing theories on inter-ethnic conflict or community erosion, the Contested Boundaries Hypothesis assumes that unclear or "fuzzy" boundaries between neighborhoods cause conflicts. Therefore, it draws attention to relationships between neighborhoods and not within neighborhoods (Dietz, 2002, 571). Legewie and Schaeffer (2016) corroborated this effect between interrelated neighborhoods with neighborhood complaints call data in New York City.

Consequently, current research concentrates on small-scale neighborhood data and the specific constellations of neighborhoods in a broader spatial context (Dean, Dong, Piekut, & Pryce, 2018). With this perspective, researchers can revise and reassess classic social science hypotheses that are still subject of an ongoing debate. Moreover, this perspective also introduces new sets of hypotheses that may have been part of theoretical considerations before but were not able to be adequately addressed. Researchers now can test more fine-grained mechanisms of social behavior or attitudes, which depicts innovation in social science research.

3.2 Common Methodological Themes

In the research sketched above, several methodological themes reoccur. Not only do researchers often face similar theoretical challenges, but they also have to deal with interconnected methodological challenges. For example, researchers, who have to find operationalizations of neighborhoods, are confronted with rather broad and also vague definitions of the theoretical concept of neighborhoods. Therefore, they sometimes vary the choice of geographic scales (Klinger et al., 2017; Tolsma & van der Meer, 2017) or they incorporate this variation in their actual research questions (Sluiter et al., 2015). The following sections analyze these questions more thoroughly as the methodological questions pose the basis for the empirical applications in this book.

3.2.1 Measurement of Neighborhoods: Ego-Hoods vs. Common Exposure Neighborhoods

In a large number of studies, the term neighborhood is not defined (Spielman & Yoo, 2009, 1100). This lack often occurs because of data limitations: many datasets only contain information on a limited number of different spatial units, such as zip codes, census tracts, district or municipality identifiers (Dietz, 2002, 541). As a consequence, researchers dismiss discussing alternative definitions for neighborhoods. At the same time, to understand and interpret the results of statistical analyses, some authors argue that it is still important to recapitulate the definition and the corresponding operationalization of neighborhoods in social studies (Sampson et al., 2002). Without knowing the implications of neighborhood definitions, inferring from data to social reality is difficult.

Chaskin (after Spielman & Yoo, 2009, 1098) differentiates three types of neighborhood operationalizations:

1. neighborhoods as social units
2. neighborhoods as spatial units
3. neighborhoods as network of associations

According to Dietz (2002, 541), neighborhoods as spatial units are the most commonly used operationalization for neighborhoods. Thus, any study on neighborhoods can automatically be understood as a spatial investigation (Spielman & Yoo, 2009, 1100). When researchers study the context of social behavior, defined by neighborhood influences, technically they analyze the variation in neighborhood characteristics on a particular outcome (i.e., social behavior or attitudes). These neighborhood characteristics are part of neighborhoods located in space. Thus, neighborhoods as spatial units impact the lives of people through their location in space *and* their associated characteristics.

At the same time, administrative units which are used as neighborhoods in a lot of multilevel studies do not necessarily predefine neighborhoods. Researchers have to define neighborhoods according to their hypotheses, and ultimately not based on what is convenient to use or is available. This effort involves searching for operationalizations of neighborhoods that represent relevant contexts of social behavior or attitudes. Georeferenced survey data are particularly useful in this effort as they allow to flexibly and dynamically model the kind of neighborhoods that are of interest.

Researchers hence use georeferenced survey data to either build ego-hood neighborhoods or common exposure neighborhoods. Ego-hood neighborhoods are individ-

ual neighborhoods for each survey respondent, for example, by drawing circular areas around their location. Tolsma and van der Meer (2017) used such areas from 500 meters up to 4000 meters to find gradual influences of ethnic neighborhood densities on social trust for which building ego-hood neighborhoods provide a particularly flexible method. Common exposure neighborhoods, on the other hand, may also be flexible in size but can serve as neighborhoods for more than one survey respondent. Researchers usually obtain them by defining borders between neighborhoods and then locating survey respondents within the borders. Förster (2018) used them to compare immigrant rates in small-scale 1 km^2 neighborhoods and administrative districts with regards to their effects on electoral turnouts. No matter which kind of neighborhoods researchers choose depending on their research question, georeferenced survey data suit both of these theoretical and methodological approaches.

Two general considerations help to decide between the two kinds of neighborhoods. First, if theories specify interaction between people according to their motion in space, ego-hood neighborhoods may be preferable. Researchers can vary the ego-hoods' size and evaluate at which geographic scale predictions of the theory apply (Sluiter et al., 2015). Second, if theories make assumptions about how specific characteristics of neighborhoods, such as policies, settlement structures, or general infrastructure, affect all people in the neighborhood, common exposure neighborhoods may be preferable. Researchers can test whether these characteristics affect all people within these neighborhoods in the same way (Thorlindsson et al., 2012). Thus, as mentioned, before researchers operationalize neighborhoods, starting with the theory does help to decide at least between these two kinds of neighborhoods.

Overall, ego-hood or common exposure neighborhoods are not necessarily distinct. Researchers can combine them, for example, by varying geographic sizes of common exposure neighborhoods. As georeferenced data are data that are relatable to other data in space, the mere number of spatial methods (see Chapter 5.2) enable such flexible approaches.

This flexibility, however, also comes with a price. Before researchers can apply these methods to their research problem, other intermediate steps and considerations are of high importance. They concern common issues in spatial research that result from the mass of information in spatial data, which affects creating measurement scales. The following section, in this regard, introduces the most prominent issue, the Modifiable Area Unit Problem (MAUP).

3.2.2 The Modifiable Area Unit Problem (MAUP)

The Modifiable Area Unit Problem (MAUP) is one of the most common issues that concerns the use of geographic data (Wu, 2007). It originates in the rationale that all areal or spatial units are arbitrary: contents of geographic scales vary and are not necessarily fixed. Depending on how researchers choose their spatial units for analysis, the results of the analysis change. "MAUP implies that results of statistical models in which contextual information is used can be strongly affected by the level at which the contextual data is aggregated" (Hillmert et al., 2017, 274), thus researchers cannot say for sure what happens to their scale in case they choose ego-hoods of 500 meters or 1000 meters. The latter is not necessarily a coarsened version of the former. Because of the MAUP, results of statistical models can differ severely and researchers have to examine their results accordingly.

Moreover, the MAUP comprises two components: the Scale Effect and the Zoning Effect (Spielman & Yoo, 2009; Wu, 2007). The Scale Effect is the variation of statistical results when spatial units are aggregated into larger units. The Zoning Effect, on the other hand, is the variation of statistical results that orginates in different methods of aggregating the spatial units. Both components may affect the results of analyses when geographic scales are varied.

Despite their conceptual difference, the impacts of the two components are similar. Regardless whether the scale of spatial units or their zoning is varied, it is unpredictable at first how they affect the results, for example, of regression or correlation coefficients. The coefficients may differ in size, in their accuracy or they even may change direction (Spielman & Yoo, 2009, 1099). One prominent case for the latter are immigrant rates in small neighborhoods in comparison to larger regions and their influence on attitudes towards immigrants: while higher immigrant rates in larger regions tend to influence attitudes positively, in small neighborhoods they affect them negatively (Putnam, 2007). Solely by inspecting the data and their distribution, researchers cannot estimate such differences of results.

Like the MAUP, other methodological challenges occur when using spatial data (Bluemke et al., 2017, 312ff). This chapter cannot discuss each of them individually, but one specific topic is exceptionally vital when researchers aim to work with individual data such as survey data. Namely, in cases in which researchers include data from different geographic levels into one analysis, or in which they try to make statements about one geographic level based on findings on the other geographic level, they have to be cautious. The following section addresses this topic in more detail.

3.2.3 Ecological Fallacies

The other prominent issue with spatial data is the risk of conducting ecological fallacies. They can occur when researchers unjustifiably transfer findings, gained on the aggregated level, to the individual level (Bluemke et al., 2017, 317). Wu (2007) used the example of illiteracy among foreign-born people in the US. Previous research on the individual level revealed that illiteracy positively correlates with foreign citizenship. On the aggregated level of states the correlation is the opposite: in regions with high rates of foreign-born people, the mean level of illiteracy is unusually low. However, the foreign-born people in these regions do not necessarily have to be literate themselves. This example shows that relationships between aggregated measures may not be the same as the relationships between individual measures.

The risk of ecological fallacies may not lead to the notion that aggregated data are inadequate to use. Wu (2007) stated misguided conceptions about the link between individual-level data and aggregated data, for instance, the assessment that "individual-level models are always better specified and more accurate than aggregated-level models, aggregate-level relationships are always intended as substitutes of individual-level relationships, and aggregate-level variables [have] no relevance to causal relationships and mechanistic explanations of individual-level activities" (Wu, 2007, 122f). Among others, variables on the aggregated level can be direct causes of individual-level processes, for example, in regards to public health issues and the spreading of diseases. It depends on the research question whether the aggregated measures correspond to the intentions of the researchers.

In the context of georeferenced survey data, researchers have to be aware that geospatial data on a small scale may represent other information than geospatial data on a higher and aggregated geographic scale. However, researchers may even intend to use this difference in their research. Förster (2018), for example, explicitly compared immigrant rates on the small neighborhood level with immigrant rates on the administrative district level. After including all these levels in one analysis, only individual-level variables remained to be statistically significant, although, as Spielman and Yoo (2009, 1102) noted, model fit may not be one of the best tools to compare models of different geographic scales. As data of different geographic scales can vary with regards to their meaning, researchers do well to refer to the theory when analyzing them.

3.2.4 Used Data

The data of studies investigating neighborhood effects are diverse because they differ with regards to their geographic scale and their actual content. This variety of data leads to a huge set of applications which, moreover, differ between countries because of unlike spatial data landscapes. Not in all countries access to geospatial data is convenient and straightforward; Germany, for example, is a country with a fragmented geospatial data landscape (Schweers et al., 2016). Multiple authorities hold data on the same topic but in different access categories and also in different geographic structures (see Chapter 4.1). Researchers who aim to use these data for their studies have to request, prepare, harmonize, and in consequence restructure these data.

Differences in geographic scales between multiple data sources often are also a consequence of disparate data availability. Not all interesting geospatial data for the use with georeferenced survey data are available on the intended geographic level; researchers have to deal with the available data. For example, researchers may want to study the neighborhood effects of unemployment rates on mental health, but these data are only accessible on the municipality level in Germany.[15] If they still want to conduct their research, they may have to adjust their hypotheses or be cautious in their interpretation of the results (see preceding section). Other data may indeed be available on neighborhood levels of 1 km^2 grid cells, such as the German Census 2011 data. From a candid standpoint, it may not be clear why the one data source is available on this small spatial scale and the other is not.

Diversity with regards to the content of geospatial data also occurs as the geosciences are a multidisciplinary field of research, alike the social sciences. Combining data from both disciplines introduces permutation that increases with each subdiscipline that gets involved. For example, geospatial data exist for traditional disciplines such as agriculture or the geography of landscapes (Plant, 2012, 9ff). There are social science data for a variety of social science disciplines, including health, education, or attitudes research. The number of possible combinations among them are extensive.

The data of the presented research examples in Chapter 3.1 often stem from the non-German context. Thus, issues of data availability may differ between the countries where the research was conducted. In Germany, for example, there is only a small amount of georeferenced survey data (see Chapter 4.1.2) which leads to a lack of a methodological institutionalization as only few researchers use them. Hence, another objective of this book is to present possible methods and applications to con-

15 https://www.regionalstatistik.de/genesis/online

tribute to the dissemination of such data.

Lastly, not all examples presented in the chapter at hand use georeferenced survey data. Some of the applications also exploited data from other sources, such as a neighborhood complaints call databases (Legewie & Schaeffer, 2016). These applications still illustrate the particular use of small-scale spatial data for social research as, in principle, such data could also be combined with survey data.

4 Challenges of Using Georeferenced Survey Data

This chapter is about the challenges of using georeferenced survey data. By reference to research using georeferenced data, the previous chapter exemplified that current applications lack common standards because of theoretical as well as methodological differences in operationalizing space and neighborhoods. Some of these missing standards indeed may be explained by still existing challenges of using georeferenced survey data. Each of these challenges—data availability, technical procedures, data protection, and privacy—hinder research from a more widespread fraction of the social science community. As long as these issues of using georeferenced survey data persist, research will lack common standards. To navigate these challenges, the following discusses them more specifically.[16]

4.1 Data Availability

Before social scientists can conduct research with georeferenced survey data, they need access to the data. One of the most important prerequisites is data availability. Moreover, as spatial linking (see Chapter 2.5 and 5.2) requires access to even two data sources, this prerequisite is twofold: it concerns the geospatial data landscape (Schweers et al., 2016), and the data landscape of georeferenced survey data.

The empirical applications in Chapter 6-8 use data from Germany, both for the geospatial data and the georeferenced survey data. Most of the remarks below, therefore, apply to the situation in Germany.[17] At the same time, the German context may also be a useful example because it depicts a rather extreme case for data availability

16 In Müller* (2019), I provided an even more in-depth analysis of the challenges of georeferenced survey data and its implications for research data management. Specifically, the challenges of applying technical procedures and data protection legislation described here (Chapter 4.2 and 4.3) are also part of Müller* (2019) which targets a more general audience interested in research data management issues with regards to georeferenced survey data.

17 Furthermore, this book concentrates on the data landscape for geospatial data from public authorities. These data have the advantage that authorities collect them according to standards which results in open and, in principle, reproducible data. Whereas commercial providers of geospatial data offer interesting geospatial data as well, their compilation of data is part of a business secret. Indeed, providers such as microm (www.microm.de) in Germany corroborate their data with other open data sources such as data from the German Socio-economic Panel (Goebel, Spieß, Witte, & Gerstenberg, 2014), yet the black-box principle of data generation remains. For convenience, and because the empirical applications in this book use publicly available geospatial data, the following discusses the data landscape for geospatial data from public authorities in Germany.

due to strict legislation for data protection and privacy. Germany is also an appropriate example because the administrative structure of the Federal Republic leads to a fragmented data landscape. Thus, while in detail the situation in other countries may involve other challenges, two important challenges of availability are already covered by the German case.

4.1.1 Geospatial Data

The geospatial data landscape in Germany is a prime example of how federal public structures affect research. Along with the general Federal Statistical Office of Germany, destatis, almost all federal states host their own statistical office (in sum 14 of 16 federal states; Hamburg and Schleswig-Holstein as well as Berlin and Brandenburg each share one office). Since a ”comprehensive supply of spatial data exists only for those policy domains where local governments are required to deliver data to the national federal office” (Schweers et al., 2016, 108), it is sometimes difficult to receive data for the whole extent of Germany in a standardized spatiotemporal structure. Geospatial data either are available for a subset of municipalities or regions, or they contain time mismatches. In any case, the geospatial data landscape in Germany is fragmented, and researchers have to navigate these issues by requesting data at different authorities and agencies.

Furthermore, this missing or decentralized supply of data also applies to data collections that public authorities collected according to European regulations. Thus, even EU obligations to collect data do not guarantee access to the data. Some of the data that the applications in this book use can serve as examples for both poles of data availability. While the German Census 2011 data are available for the whole extent of Germany at a central access point on the internet, the environmental noise data exist as an incomplete and non-harmonized data dump in the EIONET Central Data Repository. Consequently, the German Census data are far more accessible for research than the environmental noise data.

Even so, one should note that these circumstances are not the result of missing responsibilities or bad intentions. As for the examples of the German Census and the environmental noise data, their corresponding EU regulation and directive were just implemented differently in German law. The census data aim to serve as comparative data for the whole EU; the environmental noise data aim to serve as data to develop noise reduction plans for municipalities in the EU. The motives for providing access to harmonized data for the whole extent of the countries were just different.

By all means, getting access to harmonized geospatial data in Germany can merely

be answered on a case-by-case basis. Müller* et al. (2017) presented a list of comprehensive geospatial data collections in Germany which can be found in the appendix of the publication. However, this list is prone to change, either because data sources disappear, or because new ones appear. Effectively, the latter case may be even more realistic as new regulations on the EU level, specifically the directive "Infrastructure for Spatial Information in Europe" (INSPIRE), oblige more and more authorities to upload their data to a central portal at the internet.[18] To date, a fragmented but changing geospatial data landscape remains challenging for researchers.

4.1.2 Georeferenced Survey Data

It is difficult to evaluate the data landscape of georeferenced survey data in Germany exactly like the geospatial data landscape. At first, all georeferenced survey data are primarily survey data. For searching and accessing them, researchers can use well-established data catalogs or search machines, giving them access to a rich and diverse universe of surveys from which many can be downloaded directly. Relating to information on georeferencing, research and even access is not that straightforward. For example, documentation of the data can leave uncertainty with regards to the details of a survey, e.g., the geographic scale of spatial units. Strict regulations can also limit access because of data protection (see Chapter 4.3). Either way, social science survey researchers who aim to use georeferenced survey data, face a less transparent data landscape than in comparison to the regular survey data landscape.

This much is certain: the demand for georeferenced survey data as well as the number of applications using them have increased in the last years (Bluemke et al., 2017; Schweers et al., 2016). Some social science research data centers provide access to georeferenced survey data already for a long time. One prominent example is the data of the German Socio-economic Panel (SOEP) which offers geo-coordinates of panel respondents' addresses starting with the year 2000 (Goebel, 2017). Other examples are data from the German Family Panel pairfam (Schmiedeberg, 2015), the National Educational Panel Study[19] (NEPS), or data from the GGSS (Klinger, 2018). Given the already available georeferenced survey data in Germany, it cannot be inferred that no data exist.

Still, access to these data is not yet standardized. This missing standardization concerns the mode of how and where researchers can find georeferenced survey data, and it also concerns detailed instructions about the available information. Whereas

18 http://inspire-geoportal.ec.europa.eu
19 https://www.neps-data.de/de-de/datenzentrum/datenzugang/sensibleinformationen.aspx

some surveys are transparent relating to, e.g., the geospatial units researchers can use at research data centers' facilities (Goebel, 2017), others solely prepare data on request (Schmiedeberg, 2015, 3). For researchers, searching and accessing georeferenced survey data involves addressing plenty of people at the research data centers that, in principle, hold these data.

It may be argued that these issues of data availability are partly a cause of the challenges which the next two sections describe: using georeferenced data in social sciences survey research involves commitment in the sense that researchers must learn new methods and new techniques, or at least that they have to gather external expertise. Moreover, it concerns the challenges of data protection and the privacy of survey respondents. Partly, these challenges will remain prevalent in Germany, but they also lack standards which are a prerequisite for a broader data availability.

4.2 Technical Procedures

4.2.1 Geocoding

As described earlier in Chapter 2.2, georeferencing survey data requires geocoding of survey respondents' locations. Geocoding is the method of converting indirect spatial references such as the respondents' housing addresses into geo-coordinates. For this purpose, there are geocoding services that access databases with address information *and* the corresponding geo-coordinates. Accordingly, they convert this information vice versa (Zandbergen, 2014, 2). Geocoding is a reasonably automated procedure which facilitates the conversion of thousands of addresses into geo-coordinates.

Running these geocoding services as an own technological infrastructure, however, is demanding and requires up to date address and geo-coordinate databases. The actual geocoding, therefore, seldom takes place locally on research projects' computers. Instead, projects rely on the expertise of third-party providers of geocoding services such as Google, Bing, or the German Federal Agency for Cartography and Geodesy (BKG). These geocoding services differ in the quality of the geocoded addresses, but more so in privacy implications which Chapter 4.3 will discuss.

By and large, geocoding is one of the first and most essential prerequisites to create georeferenced survey data. Geo-coordinates are necessary to identify survey respondents' locations to link these locations to other geospatial information. Without geo-coordinates, researchers cannot relate locations to each other and have to rely on more traditional methods of data linking, e.g., with common identifiers (see Chapter 5.1 for more details). Geocoding concerns primarily the challenges of data protection,

but applying the method of geocoding through ready-made services is not complicated (Müller* et al., 2017).

In order to manage data with geo-coordinates and to perform spatial linking, researchers have to work with specialized software: Geographic Information Systems (GIS). While Chapter 2.1 already introduced them, the next section discusses their technical and organizational prerequisites.

4.2.2 Using GIS Procedures

GIS is a tool to manage, analyze and graphically display georeferenced data (Bluemke et al., 2017). Some of the available GIS software and service providers belong to the commercial sector, e.g., the software *ArcGIS* of the company ESRI (ESRI, 2015). Meanwhile, the open source software *QGIS* (QGIS Development Team, 2019) is a free alternative for projects with smaller fundings. For researchers in the social sciences, also the use of the statistical software *R* (R Core Team, 2019) may be interesting (Müller* et al., 2017). Given these commercial and open source GIS tools, the collection of available GIS tools is already extensive.

At the same time, using GIS procedures within the software is complex. Researchers have to learn how they can apply the procedures to their spatial linking and data analysis problem (Meyer & Bruderer Enzler, 2013, 319). Furthermore, as Chapter 2.5 signifies and as Chapter 5.2 illustrates in more detail, the amount of available methods of spatial linking is tremendous. Even if researchers use spatial linking services from third-party providers, they, at least, need training in the principles of these methods to identify differences between them. The empirical applications in this book (Chapter 6-8) may serve as an adequate example of why it is worthwhile to look at the results of spatial linking—depending on the applied methods, the results and their interpretation differ severely. Conducting research with georeferenced survey data remains an interdisciplinary undertaking that requires learning the ropes of GIS procedures.

One last issue concerning the application of GIS procedures is their computational demand. Geospatial data can be rather large in file size—they can reach gigabyte file sizes, depending on the actual dataset. Furthermore, processing and analyzing these data requires a decent central processing unit (CPU) and enough main memory. Researchers aiming to use georeferenced data have to expect an increased demand for computing power.

4.3 Data Protection and Privacy

Using small-scale georeferenced survey data requires paying attention to the complex of data protection and data privacy. Georeferenced survey data are sensitive data as explicit spatial references increase the risk of re-identifying anonymized survey respondents. Moreover, re-identifying can occur during the processing of georeferenced survey data but also during the production of statistical results from analyses. Thus, in all phases of research and data management, such as storing, geocoding, spatial linking and analyzing of the data, researchers work with sensitive data.

Indeed, even ordinary survey data may incorporate sensitive information. Explicit spatial references can intensify this problem because information about locations makes creating unique observations in a dataset more probable. Unique observations are people who are candidates for a re-identification. Moreover, the smaller the locations are, the fewer people live in these areas, which intensifies the problem even more. For example, a lawyer with seven children who lives in a known city district is easier to identify than a lawyer from whom only the country of residence is known. Because of explicit spatial references, georeferenced survey data are generally more sensitive.

The complex of data protection and privacy can be divided into two separated parts. The first part is the legal perspective on data protection which regulates how researchers have to manage these sensitive data. The second part is the perspective on the already mentioned re-identification risk and its effect on data quality. The following discusses both parts separately.

4.3.1 Legal Regulations and Spatial Linking Workflow

Most importantly, storing personal information such as addresses and geocoordinates in the same place as the actual survey results is not allowed in Germany, according to data protection legislation.[20] In common survey projects, address

20 On May 25, 2018, the General Data Protection Regulation (GDPR) of the European Union and its German implementation, the ”Law for Adaptation of Data Protection to the Regulation of the European Union” (DSAnpUG-EU, own translation) came into force. In this law, one paragraph states that personal information such as addresses and other information are allowed to be combined as long as the research and the statistical purpose require it (DSAnpUG-EU § 27 Abs.3). Thus, in some circumstances, researchers might interpret spatial linking as such a research and statistical purpose. In the former law, however, this paragraph was already included in the same wording. Moreover, still, research projects established the more conservative approach to strictly separate personal information and other information. Because of that, the following presents an approach that makes spatial

and survey information are stored in separate files or even in separate locations and can only be matched by an ID correspondence table. However, to spatially link the data—survey data and geospatial data—an indirect correspondence between the geo-coordinates and the survey data must be established. This correspondence is necessary to add information from geospatial data, e.g., dB(A) values from road traffic noise data, as attributes to the survey data using geo-coordinates. To be able to achieve this correspondence without violating the data protection legislation, researchers have to apply organizational barriers during the linking procedure.

Figure 4.1 displays how the organization of such a spatial linking procedure may look by extending Figure 2.1 in Chapter 2.3. At no time, survey information and address information are stored together. Instead, even separate storage locations hold the individual data collections: computer/server A stores the survey data; computer/server B stores the address information. Another feature is the use of a third storage point, computer/server C which stores a correspondence table that establishes a correspondence between differing identifiers stored in computer A and B. The geospatial data, on the other hand, can be stored anywhere as is signified by an asterisk.

Up to the point of linking geoinformation and survey information, processing of the data is straightforward. Using the standard tools such as *R* or *QGIS*, researchers can even spatially link the geocoded address data with the geospatial data. However, during the following steps, researchers have to apply to specific procedures and make a decision of how to deal with the geo-coordinates:

First, before linking the newly added geospatial information and the actual survey data, it should be considered to delete the geo-coordinates in the data. This step prevents a direct re-identification of survey respondents through direct spatial references. Also, relating to legal compliance, personal information and survey information are thus not combined.

Second, coarsening of the geo-coordinates may present an alternative to deleting the geo-coordinates. Aggregating the geo-coordinates to higher spatial units, e.g., city districts, reduces the issue of sensitivity of the data. At the same time, this guarantees the statistical control of spatial dependencies in later analyses (see Chapter 5.3) and supports ex-post adding of geoinformation based on these spatial units. Lastly, with regards to legal compliance, coarsened geo-coordinates no longer constitute personal information.

Besides the general organizational challenge of applying the whole workflow, geocoding may impose an issue even before the spatial linking. Geocoding services

linking possible even under such strict arrangements.

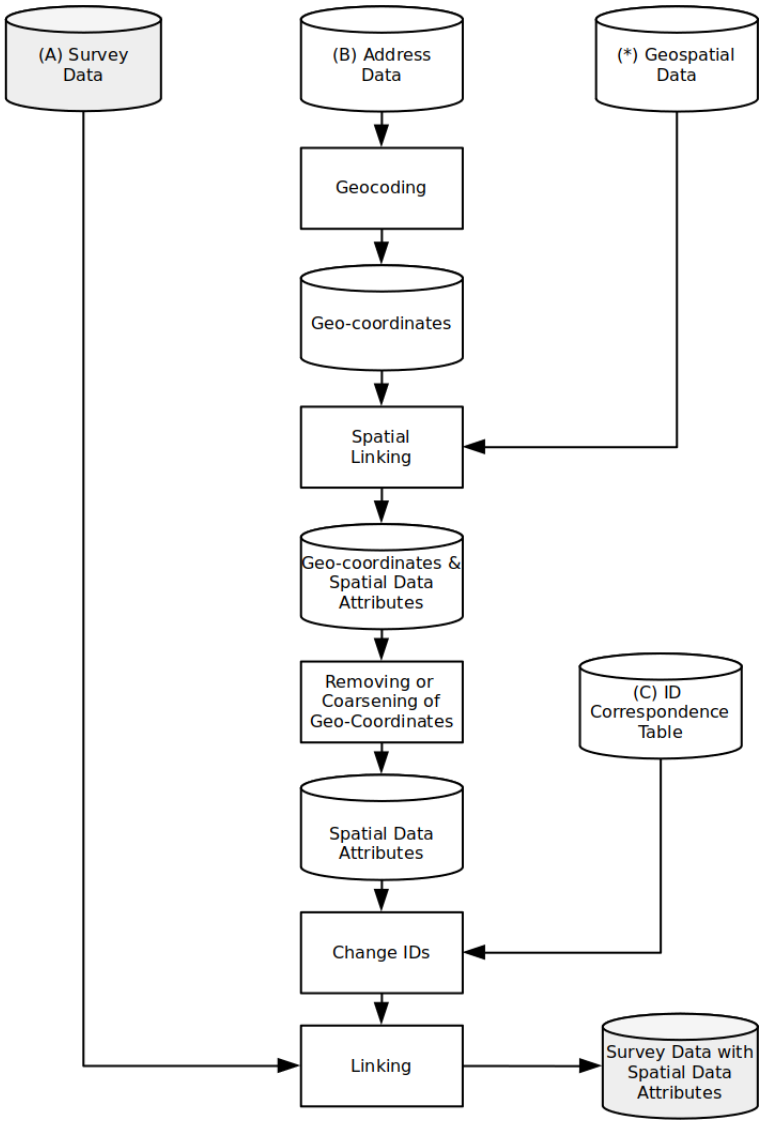


Figure 4.1: Organization of Spatial Linking Respecting Data Protection (extended)

manage data not locally on researchers' working computer. Instead, they offer the possibility to access geo-coordinates over web services for address data (Zandbergen, 2014, 2). Users of these services, for instance, upload address tables from survey respondents to the web service. After completing the conversion, they download them again, enriched with geo-coordinates. What makes geocoding services problematic is that using them implies sending personal information, which is the survey respondents' addresses, through the internet.

However, the technical implementation of the geocoding services varies. Commercial providers of these services, such as Google, might store requests for their services, containing the addresses of survey respondents. As such, relying on commercial providers may also impose the risk of non-compliance to data protection legislation. In contrast, public providers such as the German Federal Agency for Cartography and Geodesy (BKG) process requests on the fly and do not store them. Thus, they reach compliance with data protection legislation. Unfortunately, research projects at universities which are not funded by the federal state in Germany cannot necessarily use this service.[21] Researchers have to weigh up the different geocoding service providers against one another and negotiate the conditions with them to follow data protection legislation.

Even after finding geocoding services that do not store the requests permanently, e.g., by signing contracts, caution concerning their use is advised. Addresses alone may not depict personal data, but in combination with other information, addresses may help to draw further inferences about survey respondents' identity. This information may be meta information, such as a survey's project title in the filename of the address list sent to the geocoding service, or a correspondence between the rows of the address list and the response number in the survey. Furthermore, IP addresses may deliver information about the researcher's location to the geocoding service, which can provide knowledge about the specific research project. To sum up, researchers have to pay attention to many details when managing geo-referenced survey data.

4.3.2 Re-Identification Risk and Data Quality

The second part of data protection and data privacy concerns the enrichment of survey data with new information. Locating people in space on a small scale increases

21 An alternative would be to build an own geocoding service. For this purpose, for example, data from the free map project OpenStreetMap is useful, but curating these services is difficult. Also, geocoding based on exact addresses containing house numbers is not always possible since not for all addresses house numbers are available.

the risk of re-identifying survey respondents. As the number of respondents in surveys decreases with decreasing sizes of spatial units, it is easier to identify individual people with specific characteristics. Knowing survey respondents' locations in combination with their survey answers may disclose more information on their identity as intended.

Moreover, irrespective of specific locations, the locations' attributes can also produce unique observations in the data. For example, the exact number of unemployment rates or the amount of air pollution in a specific neighborhood can reveal its location, even if the name or identifier of the neighborhood is unknown. In combination with other information, such as basic sociodemographic characteristics, inferences about single persons in the corresponding survey data may become probable. Thus, the information on locations alone carries the risk of re-identifying survey respondents.

If both data sources, survey data and geospatial data, are machine-actionable, this risk even increases. Not the insider knowledge of people who gather additional information about neighborhoods can provide additional attack vectors, but the data themselves can be systematically searched and processed by potential attackers. The freely accessible geospatial data which are used for spatial linking purposes of survey data can be accessed by anyone. Researchers can lose control over this process, such that machine-actionable and free data produce another hazard for re-identifying survey respondents.

In order to decrease this risk of re-identifying survey respondents, different methods are discussed. Generally, these are methods that either veil or coarsen geo-coordinates to wipe out unique cases in data. In the first group of methods, permutation procedures are applied (Kroll & Schnell, 2016; Zandbergen, 2014) which aim to prevent any reference to the original geo-coordinates. In the second group, aggregation procedures are used to assess the risk of re-identifying people in spatial units, such as municipalities or zip code areas (Blatt, 2012; El Emam, 2006). The application of both methods is complex, depending on the structure and the content of the geospatial information, they differ severely. For this reason, they will not be discussed in more detail.

To summarize this section, at any time during research with georeferenced survey data, data protection and data privacy are essential. Starting from geocoding and spatial linking to the analysis of georeferenced survey data, researchers have to protect sensitive information. Because of sensitivity, providers which offer these data for secondary use in Germany, such as SOEP, NEPS or pairfam solely give access to their data in on-site working environments. The same holds for the data in this book which are

available at the GESIS Secure Data Center in Cologne, Germany. Controlling the input and output of data in these environments minimizes the risk of re-identifying survey respondents and allows their use to a broader research audience.

5 Linking Spatial Information to Georeferenced Survey Data

Thus far, the previous chapters prepared the terminology and landscape for spatial analysis in the social sciences. The following chapter presents how spatial methods can help to gather new information from geospatial data for georeferenced survey data, and how these combined data can be analyzed. This chapter introduces the methods that are deployed in the empirical applications of Chapter 6-8.

5.1 Difference to Other Approaches of Data Linking

Data linking means to combine or establish a correspondence between two or more distinct data sources. Common identifiers or structural similarities of the data sources enable the linking so that researchers can draw on one single file for their analysis. This approach allows to either add new attributes or to add new observations to existing survey data. Data linking in the social sciences consequently subsumes a family of methods to enrich existing survey data.[22]

As a common approach, researchers use data linking methods to add attributes of a broader geographic context to their data. If they have access to identifiers such as municipality codes, they can use external data sources from statistical offices to link them with their survey data column-wise based on common identifiers (Hillmert et al., 2017, 272). The idea is to provide auxiliary explanatory variables to the data and to test theories on the broader geographic context of social behavior or attitudes.

Alternatively, researchers use data linking methods to add new observations from other surveys to their data. For this approach, they use sets of harmonized variables between these two data sources (Zhang, 2012, 51). If the two data sources are structurally equal, they can be linked by merely combining them row-wise. The idea is to

22 In other scientific disciplines, the term data linking can also mean to link two data sources on an ontological level. The two data sources then are not linked physically. Instead, correspondence is established by finding common identifiable information (Hallo, Luján-Mora, Maté, & Trujillo, 2016). Use cases comprise, for example, finding cited data sources in a research publication and then connect both, the data and the publication, ontologically based on this identifiable information (Bensmann, Zapilko, & Mayr, 2017). On the long run, by connecting different sources, a network of publications and data evolves that can be analyzed with regards to the citing practice of researchers. This book, however, leaves aside this understanding of data linking. The empirical social sciences generally use data linking to enrich their social sciences data with auxiliary information for their analyses.

increase sample sizes for analyses and gain more reliable statistical results.

Figure 5.1 displays these two approaches. Column-wise linking combines the two data sources A and B by using common identifiers. Row-wise linking combines the two data source by establishing structural equivalence through harmonizing A and B. Also, using more than two data sources, or even mixing the approaches with at least three data sources, is possible. For instance, researchers could combine two data sources and subsequently add attributes from a database of the statistical offices based on some common identifiers, such as municipality codes. Linking data sources column-wise and row-wise opens up plenty of possibilities for research.

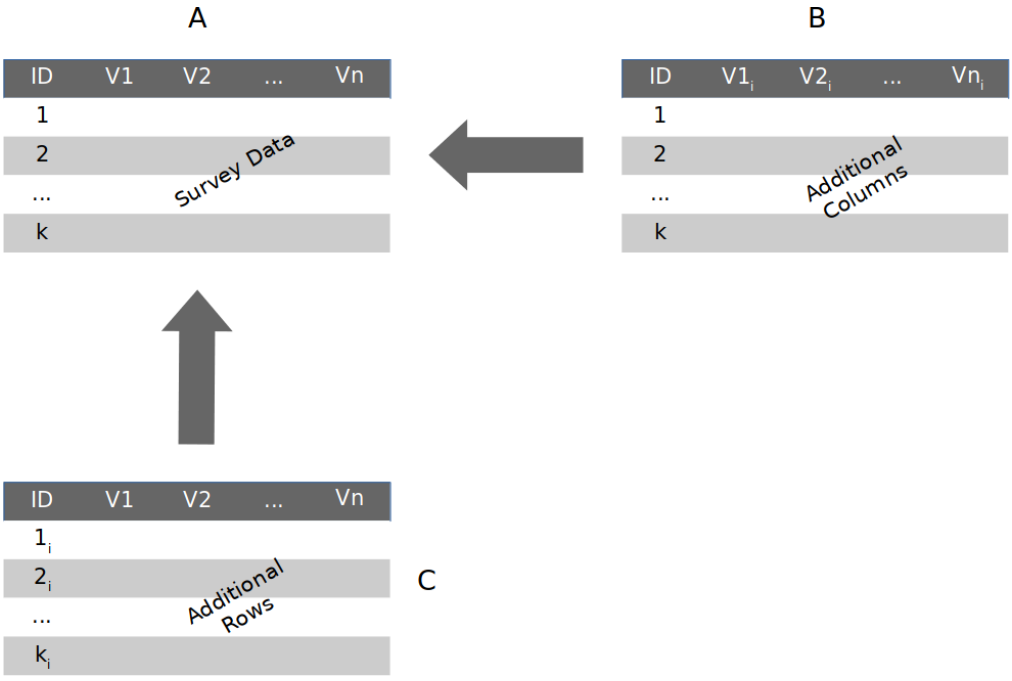


Figure 5.1: Column-Wise and Row-Wise Linking of Additional Data to Existing Survey Data

Another specific case of these data linking efforts is record linkage. This method aims to link data from different data sources that contain individual data, i.e., data from single persons which were collected separately and for different purposes. Prominent applications are the linking of medical records to create a better understanding of the development of diseases (Taylor, Irvine, Iannotti, Harchak, & Lim, 2014). Such studies address statistical methods to combine the data, for example, either by finding deterministic or probabilistic links between them (X. Li & Shen, 2013, 33ff). However, legal provisions and survey response issues complicate record linkage with regards to

the consent of respondents (Sakshaug, Couper, Ofstedal, & Weir, 2012). Accordingly, record linkage involves extra challenges regarding the statistical and organizational methods of linking compared to the above presented classic approaches of data linking.

How does spatial linking fit into these data linking approaches? Spatial linking belongs to the column-wise approach of data linking. Its goal is similar—adding new attributes to data—, but the linking does not require common identifiers. Instead, in this case, geo-coordinates make it possible to project both data sources into a joined geographic coordinate space. The projection, in turn, establishes a geographic relationship between each of the geo-coordinates. As such, spatial linking does not require linking two data sources one-to-one based on strict identifiers, and is a more flexible version of column-wise linking.

Thus, relating two data sources with spatial linking is even possible for observations that are far apart. Geo-coordinates can be used, for example, to calculate geographic distances between relevant points of the two data sources. These geographic distances may be relevant information on their own but even more so for advanced calculations such as summary statistics or other descriptive statistics. Finding common identifiers is not necessary for spatial linking—drawing on geo-coordinates that are part of the same CRS is sufficient.

5.2 Types of Spatial Linking

Combining spatial information from different sources through spatial linking comprises a set of highly standardized GIS techniques. What is similar between the different techniques is that most of them make use of topological models. These models ensure that any calculation or transformation of the data preserves the association between separate objects or observations in the data (Zhong, Jing, Chen, & Wu, 2004). In consequence, no matter which technique of spatial linking is chosen, the geographic associations between geometries remain to be intact. Calculating measures based on geographic distances, for example, is not different comparing a radius of 100 meters and another of 1000 meters. The base of any spatial linking technique does not change when changing the parameters of the calculation.

Some of the spatial linking techniques in this section refer to the *Dimensionally Extended Nine-Intersection Model* (DE-9IM) (Strobl, 2017). The DE-9IM is a topological model for vector data to establish spatial relationships between two datasets and is

incorporated in widely used GIS standards, such as the ISO 19125[23] or the standards of the Open Geospatial Consortium (OGC) (Percivall, 2016). This topological model is also known as the *Simple Features Interface Specification* or briefly *Simple Features*. GIS software, moreover, such as *QGIS*[24] or *R*[25], rely on this standard for some of their spatial linking methods. Accordingly, the DE-9IM model also provides an adequate tool for social science researchers who aim to conduct spatial linking methods on georeferenced survey data.

The DE-9IM contains methods to identify intersections between two spatial vector geometries. Definitions of boundaries between them, as well as interiors and exteriors of objects, enable a Boolean logic in the form of a three × three matrix—hence, the name nine-Intersection Model. Furthermore, this model enables exploiting topological predicates to specify the character of intersections. For example, the predicates can specify whether one spatial object contains another spatial object or whether they cross each other. Thus, the DE-9IM inherits an extensive range of applications for an extensive range of research questions.

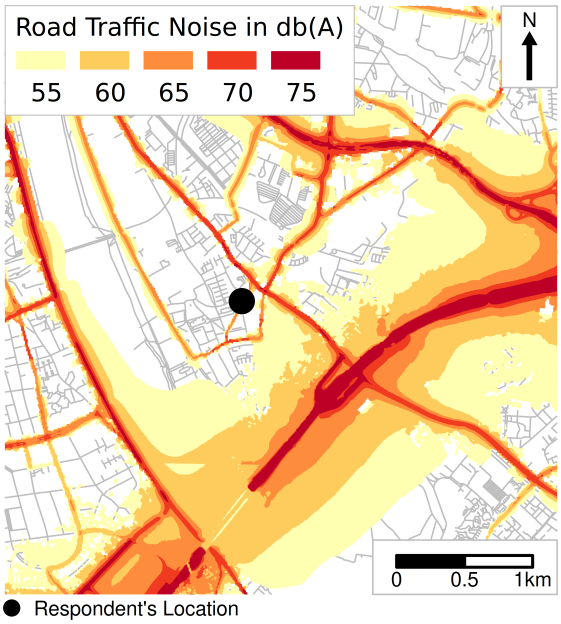
Moreover, specifying intersections by using topological predicates is possible along different vector geometries, such as polygons or lines. Georeferenced survey data based on the respondents' addresses, however, contain point data of respondents' locations. For example, intersections between these points and polygons may be of interest, whereas intersections between polygons and and polygons are not. Because social science survey researchers cannot use all of the DE-9IM procedures, the following sections exemplify a specific collection of spatial linking methods which the empirical applications in Chapter 6-8 exploit.

These spatial linking methods also contain methods to combine point data with raster data that are not part of the DE-9IM model. Raster data comprise evenly shaped raster cells for a specific geographic area. Methods for calculating spatial relationships or other statistics make use of this plain data structure by applying matrix operations for descriptive statistics. All raster cells have the same geographic extent, and they solely differ in their corresponding attribute value. Accordingly, these spatial linking methods with raster data are based on mathematical calculations using two-dimensional data matrices with corresponding attribute values and not topological models.

23 https://www.iso.org/standard/40114.html
24 https://docs.qgis.org/testing/en/docs/pyqgis_developer_cookbook/geometry.html
25 https://www.r-consortium.org/blog/2017/01/03/simple-features-now-on-cran

5.2.1 Linking by Location

The simplest method of spatial linking is linking by location, a method that links two data sources one-to-one. It transfers a value at a specific location of one data source to the other data source at the corresponding location. Comparable to linking approaches with common identifiers, the transferred values are the same in both data sources. Moreover, what is different from linking with common identifiers such as zip or municipality codes is that, in practice, spatial linking by location also works on a smaller geographic scale. Geo-coordinates replace the search for harmonized identifiers, and as long as they match, spatial linking of two data sources is straightforward.

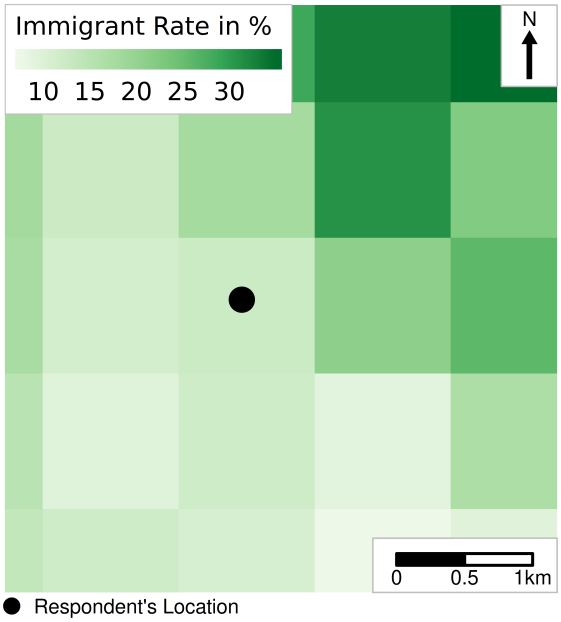


Data Sources: German Environmental Agency / EIONET Central Data Repository (2016) and OpenStreetMap / GEOFABRIK (2018)

Figure 5.2: Spatial Linking by Location with Road Traffic Noise Data

Figure 5.2 displays this method of spatial linking at the example of road traffic noise. The map projects the fictional geo-coordinates of a respondent in the same coordinate space as road traffic noise data. Spatial linking by location allows assigning the dB(A) value of the road traffic noise to the survey respondent's location. After applying the data protection arrangements, described in Chapter 4.3, the results of this method of spatial linking are extra variables in the survey data that depict the dB(A) values measured at the respondent's locations. Another example of spatial linking by location is the rate of immigrants in Figure 5.3. The method transfers the percentage of immi-

grants in 1 km^2 raster cells to the respondent's location which happens to fall into the corresponding raster cell. As a result, the survey data comprise one extra variable that depicts the immigrant rate at the corresponding location.



Data Source: Statistical Offices of the Federation and the Länder (2016)

Figure 5.3: Spatial Linking by Location with Immigrant Rates Data

This method of spatial linking yet is prone to inaccuracy errors for the following reasons:

First, geospatial data can be inaccurate through measurement errors or missing data (Henry & Boscoe, 2008). While this may not pose a problem for statistical models based on the geospatial data as such, the transferred values, as part of the survey data, are nothing but a subset of the geospatial data values. Thus, there may not be enough data points to compensate for measurement errors.

Second, inaccuracies can occur because of the geographic scale of the geospatial data. If their values are a product of aggregating data from a smaller spatial unit to a higher spatial unit, the variance on the original scale is reduced (Bluemke et al., 2017, 315). A useful example is the German Census 2011 data as the raster cell values are mean values within a sizable geographic area of 1 km^2. All the variance within these areas is assumed to be zero.

Third, also the geo-coordinates of the survey respondents' locations can be inaccurate. Depending on the service, geocoding of addresses often results in centroid geo-

coordinates of the building at the corresponding address. However, measurements of road traffic noise, for example, stop at buildings' facades. Spatial linking by location assigns a value of zero to the respondents, although it is unrealistic to assume that respondents cannot hear any noise in the building. Another example concerns respondents who live at the border of two census 1 km^2 raster cells—which one has the higher influence on the respondents? Accordingly, researchers have to consider potential biases when they interpret the results of spatial linking by location.

Lastly, on a different note, it may not be appropriate to concentrate on single location values for specific research questions. While actual noise measured at respondents' buildings may be a determinant of health (Babisch et al., 2014), its associated factors such as odor or vibrations influence people's well-being as well (Oiamo et al., 2015). Measures of surrounding noise sources can also be of interest because they affect people through these associated factors. In these cases, not the noise pressure level in dB(A) has an effect, as it is too far away, but the odor or vibrations. Likewise, not merely the social context within 1 km^2 grid cells is relevant for people. They commute to work, they shop in city centers, and they make visits to other people. Accordingly, what people experience in these locations affects them as well. Using spatial linking by location cannot answer complex questions on social behavior and attitudes that involve different characteristics of space or movement in space.

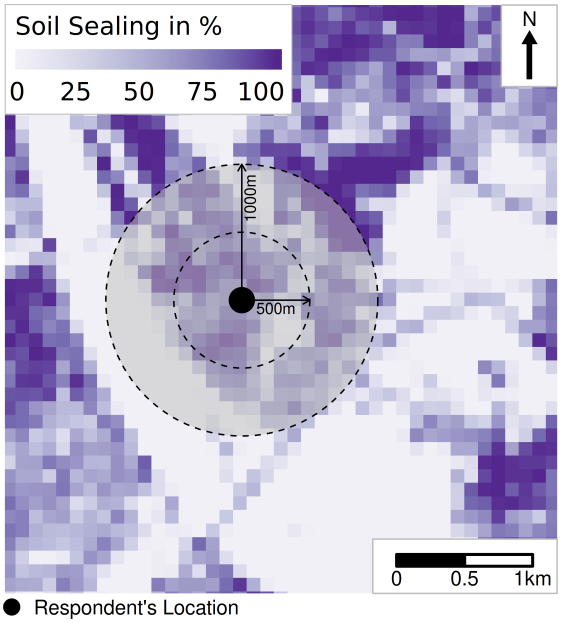
Georeferenced data enable more methods of spatial linking by projecting data in a joined coordinate space. Projection allows relating different observations, not just one-to-one but also through proximity. The following section presents methods that exceed spatial linking by location.

5.2.2 Buffers

A commonly used method of spatial linking is to build buffers as the empirical application in Chapter 8 demonstrates. Buffers are circular areas around certain geometries that can vary in size (De Smith, Goodchild, & Longley, 2018). They are used to calculate descriptive statistics of the area around each corresponding point. Information of surrounding spatial objects is combined to calculate, e.g., mean values, maximum values, or standard deviations. As the size of the buffers can be varied, they provide a flexible tool to assess growing or shrinking influences of surrounding neighborhoods (Sluiter et al., 2015; Tolsma & van der Meer, 2017).

Buffering is often used to combine spatial points data—i.e., survey respondents' geocoordinates—with raster data. All raster cells that fall into a buffer form the basis to calculate the corresponding descriptive statistics. Moreover, they are not weighted,

neither by the geographic distance to the geo-coordinate nor by the area that they extend. Buffer calculation makes use of a relatively plain data structure, which makes them easy to deploy on rather large datasets.



Data Source: Leibniz Institute of Ecological Urban and Regional Development (2018)

Figure 5.4: Spatial Linking by Buffers with Soil Sealing Data

Figure 5.4 displays the calculation of buffers at the example of soil sealing. Circular areas of 500 meters and 1000 meters around a specific geo-coordinate are defined. All raster cells that fall in the corresponding circle are used to calculate the matching statistics. In evenly distributed areas which have low levels of spatial clustering, the size of the actual buffers does not make a difference. If the aimed attribute, however, is unevenly distributed, different sizes of buffers can make a considerable difference. Soil sealing in high population density areas, for example, is more evenly distributed and has low levels of spatial clustering.

Using buffers enables researchers to incorporate information from surrounding areas and not just the specific location of a survey respondent. This method is interesting because it is unrealistic that solely direct neighborhood attributes influence the attitudes and the behavior of people—the action radius of people is bigger. Furthermore, researchers can vary the sizes of the surrounding areas. Sluiter et al. (2015) varied the sizes of buffers to find the most influential size of neighborhoods regarding ethnic diversity and its effect on social trust. While researchers have to be cautious to assess

neighborhood effects not exclusively based on model fit (Spielman & Yoo, 2009), this exemplifies the flexibility of buffering methods. They allow testing different competing hypotheses on neighborhood influences on people's behavior and attitudes.

Again, two ideas should be considered with regards to the use of buffers: First, buffers operationalize neighborhoods as circularly formed ego-hoods. Natural boundaries, such as rivers, cut these circular areas off the surrounding territories and make it hard to overcome them physically. In cases where buffer areas depict areas of social interaction within a neighborhood, large streets can obstruct interaction. For these reasons, it is questionable whether people perceive their neighborhoods as being circular. Characteristics of the physical environment of people could introduce uncertainty in the measurement of buffers.

Second, buffers entail information on surrounding areas of a focal point. Some hypotheses in social science research though also involve assumptions about the associations *between* neighborhoods. These relationships occur on an aggregate-level, yet they also affect people on the individual level. Legewie and Schaeffer (2016), for example, showed that contested boundaries of ethnicity between neighborhoods increased people's complaints about neighborhood behavior. Thus, the constellation of neighborhoods in a network of related neighborhoods influenced people's behavior. Accordingly, researchers cannot use buffers alone to operationalize complex relationships between neighborhoods or ego-hoods.

The first consideration is challenging to address. It requires complicated GIS procedures to filter for natural boundaries and to compute conditional buffers based on these constraints. Even then, it remains difficult to assume that the same type of boundary affects different persons in the same way. Seniors, for instance, may have a harder time to get from one point within a buffer to another, while younger people may not. At the same time, it still can be argued that buffers depict at least a useful approximation of neighborhoods that are worthwhile to inspect. Researchers should merely remain cautious with regards to the issues of natural boundaries or differences in the perception of neighborhoods when interpreting the actual results of statistical analyses.

In order to address the second consideration, other spatial linking procedures are available. They allow to weigh neighborhoods according to their proximity to a focal point, but also to relate surrounding neighborhoods to each other. The following section presents one of these procedures.

5.2.3 Focal Linking

A method to relate neighborhoods to each other is focal linking. Comparable to buffers, it makes use of uniformly shaped raster data. Instead of capturing all information in a specific radius, focal linking uses schemes of grid cells weighted via matrices around focal grid cells (Raju, 2004, 168)—hence, the name focal linking. Figure 5.5 shows such matrices that relate information of one focal grid cell to other surrounding grid cells. The raster cells depict operationalizations of related neighborhoods on an aggregate-level. Comparable to buffer calculations, the matrices in this method can be used to calculate (weighted) descriptive statistics of neighborhoods.

$$\begin{matrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{matrix}$$

(a) Equal weights

$$\begin{matrix} .5 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & .5 \end{matrix}$$

(b) Different weights

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix}$$

(c) Focal cell only

$$\begin{matrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{matrix}$$

(d) Focal cell left out

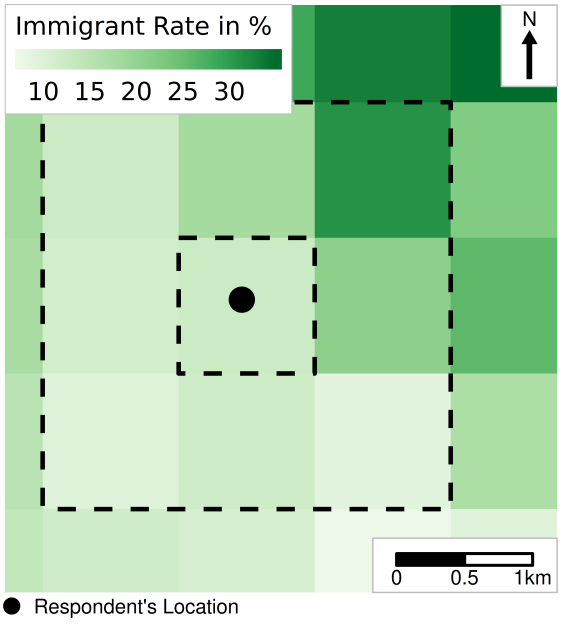
$$\begin{matrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{matrix} \quad - \text{ or } \div \text{ or } \times \quad \begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{matrix}$$

(e) Relating matrices to each other using mathematical operators

Figure 5.5: Variations of 3×3 Focal Neighborhood Matrices

The matrices can represent a multitude of neighborhood relations (Figure 5.5 a-d). Moreover, by purposely excluding the value of the focal raster cell in which a survey respondent lives, researchers can gather information from surrounding neighborhoods solely (Figure 5.5 d). In turn, they can relate this information to information from the focal grid cell—the direct neighborhood—, e.g., by building ratios, difference scores or interactions (Figure 5.5 e). Distinct from buffers, these relations do not necessarily comprise descriptive statistics of all surrounding areas of a focal point. Moreover, they also contain a potential variance between direct and surrounding neighborhoods. Focal linking allows answering more fine-grained and detailed research questions about neighborhood effects.

One example of such constellations, which the empirical application in Chapter 7 of this book uses, is based on data on immigrant rates from the German Census 2011. The raster cell, containing a survey respondent's geo-coordinate, depicts the focal raster cell and the direct neighborhood. Surrounding grid cells describe the surrounding neighborhoods accordingly, as displayed in Figure 5.6. Both sets of information can be used to compare them to each other, for example, by dividing them to build a ratio, subtracting them to build a difference score or to multiply them to include them as an interaction term in a statistical regression model. Either way, each of these measures serves as operationalization to answer social sciences research hypotheses, e.g., with regards to their effect on attitudes towards immigrants (see Chapter 7).



Data Source: Statistical Offices of the Federation and the Länder (2016)

Figure 5.6: Spatial Linking of Immigrant Rates Data with Focal Matrices

Focal linking is an alternative to buffers if the size of the used raster grid cells is large, as in the case of the German Census 2011 data which contain raster grid cells of 1 km^2. Calculating buffers of different sizes does not make sense in such cases. Adjacent grid cells of a focal grid cell could be the ones that are relevant. Buffers of different sizes, however, are helpful if grid cells are small. An example is the raster grid cells of the soil sealing data which depict an area of 100 meters $\times$ 100 meters. Nonetheless, another advantage of focal linking is the fact that surrounding grid cells can be weighted according to their distance to the focal point. Thus, while buffers and focal linking

make use of the same data structure, the results of their application differ.

Both focal linking and buffers are suited for geospatial data that cover a rather broad extent of an area. In cases of systematically missing data, for example, because of censoring, focal linking and buffers may capture too little information. Accordingly, these methods would be inefficient to use. Alternatives to these GIS procedures, however, exist that are better suited for such cases. One of these methods is the calculation of distances between geometries, which is qualified to navigate issues of non-extensive geospatial data and, moreover, to answer even more sets of research questions.

5.2.4 Geodesic Distances

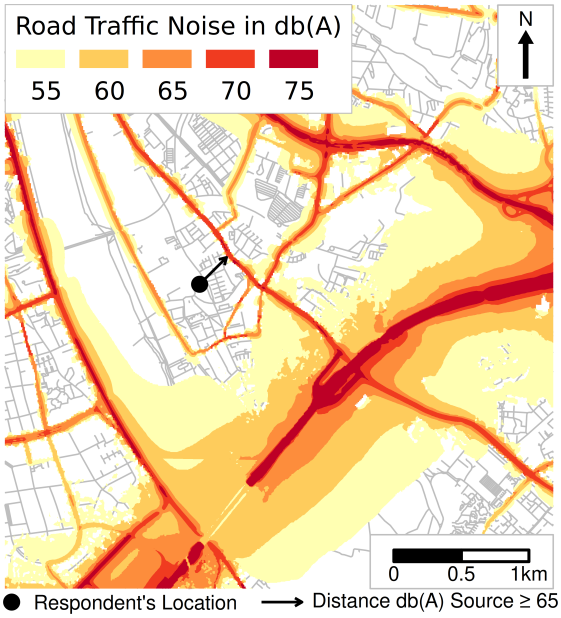
Not all geospatial data that are interesting for social science research cover a broad geographic extent. Some data are available as single points data or as polygons which are not in the direct surrounding of survey respondents. To spatially link them by location, buffers and focal linking are not useful because both of them would result in a considerable amount of non-matching cases or inefficient solutions (see the previous section). For these reasons, the spatial linking of such data with georeferenced survey data requires other methods.

One of these methods is the calculation of geodesic distances, used in Chapter 6, which produces distances between two geo-coordinate pairs. Geodesic distances make use of the projection of geo-coordinates on earth's surface (see section 2.4.1) and differ from calculating a straight line between two points on a flat surface; instead, these distances are the beeline between them that also respect earth's curvature. Geodesic distances represent a commonly used method in research.

Moreover, geodesic distances display the enormous potential of GIS procedures and the associated projection in a common coordinate space. Geometries cannot only produce relationships within short ranges but also between distant ones. In principle, links between two points which are on the opposite sides of the globe can be established. Geodesic distances provide a flexible tool to model short range and long distance relations between two geometries.

An example of the use of this method is the proximity of survey respondents to amenities in neighborhoods, such as kindergartens. In this case, minimal geodesic distances to the next kindergarten can serve as an appropriate measure of proximity. The calculation of geodesic distances then requires two sets of geo-coordinates: the geo-coordinates of the survey respondents' addresses, and the geo-coordinates of the kindergartens. After applying the geodesic distance calculation, the results illustrate minimal distances between these sets of geo-coordinates and depict how close

respondents live to the next kindergarten. Researchers can use this information to analyze whether this proximity influences people's perception of how they can combine family and career goals.



Data Sources: German Environmental Agency / EIONET Central Data Repository (2016) and OpenStreetMap / GEOFABRIK (2018)

Figure 5.7: Spatial Linking by Distances with Road Traffic Noise Data

Calculating geodesic distances, however, is not limited to geometries of the same kind. Geodesic distances between point locations and, e.g., the edges of polygon geometries can be computed as well. Researchers can use this approach to measure geodesic distances between a survey respondent's house and the next park. Moreover, they can filter geometries of interest by specific criteria or attribute values of other geometries. Figure 5.7 displays an example of a point geometry and its geodesic distance to the next road traffic noise source of at least 65 dB(A). Geodesic distances are useful to approximate neighborhood features that are not directly observable as in the case of spatial linking by location. Accordingly, geodesic distances provide another flexible tool of spatial linking for a broad range of different applications.

5.2.5 Collection of GIS Procedures

The presented methods of spatial linking are not an exhaustive list of all available methods. Other methods in GIS offer different options to manipulate georeferenced data: they can be combined, interlinked, and processed in multiple steps. Focal linking, for example, is a method that combines two procedures in one step: the conversion of raster data through the neighborhood matrices and spatial linking by location for the focal raster cell. Accordingly, in other applications, other combinations of GIS procedures are used that cannot be presented here.

Table 5.1: Spatial Linking Methods and Their Use in the Empirical Applications (I-III)

Application	By Location	Buffer	Focal Linking	Geodesic Distances
I	✓			✓
II	✓		✓	
III	✓	✓		

The presented methods are those that are used in the empirical applications in Chapters 6-8. Table 5.1 gives an overview of the methods across the individual applications. In any application, spatial linking by location is applied, either because it is the method of choice or because extra information is added with this method, such as municipality level inhabitant sizes. Buffers are only used in Application III (Chapter 8), focal linking in Application II (Chapter 7), and geodesic distances in Application I (Chapter 6).

5.3 Statistical Models for Analyzing Georeferenced Survey Data

Georeferenced survey data are not necessarily different from other social sciences survey data. If researchers use them to add spatial information to their data, as described above, the only difference is that they are enriched with additional variables. Still, the content of these variables can differ from information gathered in standardized surveys. For example, dB(A) noise values represent measurements on a logarithmic scale. While logarithmic scales are not unusual in the social sciences, researchers have to know how to interpret them in the particular use case of noise measurement. What are unusually high values that affect people's health, and what are moderate values? Thus, using georeferenced survey data requires interdisciplinary efforts.

Apart from this content-related component of new variables, researchers may also conduct methodological considerations when analyzing the data in statistical models.

One of the most prominent considerations affects spatial dependency which implies the idea that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, 236), a phrase that is also known as *Tobler's First Law of Geography*. This law states that people who live close to each other are more similar concerning specific characteristics, such as sociodemographics or attitudes, than people who live far away from each other. This similarity in turn also implies conformity of observations in data. Applying common regression models such as *Ordinary Least Squares* (OLS) would violate assumptions of observations' independence. Comparable to the approaches of other dependency frameworks for survey data, such as in multilevel regression models, researchers have to address dependencies in geospatial data attributes when analyzing the data.

Moreover, another problem is caused by the endogeneity of attributes. Using attributes from neighborhoods often means using aggregated data which are the result of social interaction or the characteristics of neighborhoods' inhabitants. If researchers aim to link this information to survey data, they link data that stem from the same people. Thus, in the study of neighborhood effects, the neighborhood attributes sometimes are no longer exogenous. Also, people choose their neighborhood; they decide where they want to live (Dietz, 2002, 551). Accordingly, moving is carried out by the people themselves and is not exogenous either. Sources of endogeneity are often as manifold as the people who live in the neighborhood because attributes of neighborhoods often directly stem from these people.

The problem with endogeneity is that non-exogenous covariates in regression models produce severe estimation biases. Endogeneity occurs in statistical models when one or more predictor variable correlates with the disturbance term of a regression model (Wooldridge, 2012, 513). As disturbance terms aim to capture all variance that exogenous variables do not explain, correlating disturbance terms and predictor variables imply a model design issue. In consequence, estimates are unreliably biased, and the coefficients of the model cannot be trusted.

In sum, researchers face severe estimation issues when they use common statistical models, such as OLS regression, in their study of neighborhood effects. Spatial similarities of survey respondents can result in the issue of observational dependencies, endogeneity and, moreover, *Omitted Variable Bias* (OVB) (Dietz, 2002; LeSage & Pace, 2009, 550f). As a large number of studies in the past used such common statistical models, in his seminal paper on neighborhood effects, Dietz even came to the rather harsh conclusion "that existing studies of neighborhood effects that employ OLS estimation are of little scientific value" (Dietz, 2002, 551). Against this background, approaching the issue itself as well as potential solutions to it seems reasonable.

The following section first presents a set of commonly used regression models to analyze georeferenced survey data. It will then turn to econometric models that formulate spatial dependencies in data explicitly and that are considered as more adequate for the above-presented issues. Moreover, these econometric models are useful for research questions in which measures of spatial diffusion are of interest. Lastly, a concluding section discusses further avenues and the implications for the empirical applications in the following chapters.

5.3.1 Commonly Used Regression Models

To analyze multivariate relationships in data, one of the most straightforward approaches is to use OLS estimation. Apart from its well-known specification, one of its strictest assumptions is that observations' error terms in the data are independent of each other (Gelman & Hill, 2007, 46). In practice, this assumption implies that the outcome y_i of one observation i does not affect the outcome y_j of any other observation j. All relationships that researchers find in data result from the relationship of variables within observations and not between observations.

The assumption of independence often does not hold (Wooldridge, 2012, 101). In survey data[26], respondents from one geographic cluster A tend to give similar answers as other respondents living in cluster A compared to respondents from cluster B (Gelman & Hill, 2007, 237f). The same effect exists in a reverse direction.

As mentioned, dependent observations in data severely affect estimates of regression models. One of the most common issues is that standard errors tend to be underestimated (Angrist & Pischke, 2009, 231ff). Researchers who ignore this bias, therefore, increase their chance of conducting type I statistical errors (Cohen, 1994, 1000). They risk rejecting the null hypothesis while in reality, it may be more appropriate to assume it is true. In other words: Not adjusting for the independence of observations can result in statistical relationships that do not exist.

Because of this issue, alternative approaches to OLS regression aim to control or adjust the clustering of observations, for example, by using dummy variables for geographic clusters (Wooldridge, 2012, 488ff). This approach can turn into inefficient modeling as the number of clusters increases. While in models with a low number of clusters using dummy variables can pose a convenient approach, models with high numbers of clusters can again be difficult to estimate. In cases in which the number of clusters k approaches the number of observations n, reliable estimates are no longer

26 Given the assumption that a survey follows a clustered design in which first, a sample of clusters are drawn and second, people within these clusters are chosen.

possible to realize. As useful as such models are because of their direct applicability, they become less suitable for more complex data structures with regards to reliable estimates.

Ultimately, using the dummy variable approach for geographic clusters also leaves no room for evaluating the actual influence of the clustering on the results. Such models do not provide any information on the variance of the outcome within and between the geographic clusters (Gelman & Hill, 2007, 252ff). This variance is often crucial in specific research settings in which researchers analyze potential differences between geographic clusters. For example, researchers may study whether two or more different city districts (i.e., geographic clusters) differ with regards to their sociodemographic composition. Comparing only the coefficients of a large set of dummies would not support them in this regard. Thus, to analyze the influence of geographic clusters on estimates researchers have to seek alternatives to dummy variable approaches.

A standard option to decrease standard error biases and to grasp group statistics of variances are *Multilevel* (ML) models (Dietz, 2002, 555f). ML models provide accurate standard errors for estimates, and they also allow estimating variances of parameters on the observation level across clusters. Researchers can estimate such variances for intercept parameters to account for varying levels of the outcome across clusters. Moreover, they can estimate variances for slope parameters to assess varying covariate effects on the outcome across clusters (see, e.g., Förster, 2018). When geographic clusters are of specific research interest, ML models provide extensive information for researchers.

This extensive information comes with the prize of requiring additional effort in interpreting and checking the models. ML's coefficients are computed under strong statistical assumptions as well, which means that researchers have to examine the data generation process in particular. For example, the estimation of the variance components assumes that clusters stem from a random sample, leading to a normal distribution assumption for the variances. While this specific assumption can be relaxed by using alternative but more complex estimators (Gelman, 2006), what remains is that researchers still have additional information to deal with such as intra-class correlation coefficients or statistical significance tests of variance terms. Researchers who switch from simple OLS regression models to more complex ML modeling to control spatial dependence, at the same time, invest in another set of ambitious modeling assumptions.

However, if content-related differences between geographic clusters are not of particular interest, *Fixed Effects* (FE) regression models may provide another alternative to OLS regression models (Dietz, 2002, 554). These models allow controlling between-

cluster heterogeneity while also providing efficient estimators. Estimated coefficients of FE models are within-cluster estimates stemming from the elimination of all unobserved heterogeneity between clusters. Often used for panel data (Wooldridge, 2012, 484ff), previous research has shown that these estimators are also well suited for geographically clustered data, for example, as they also tackle issues of OVB (Dietz, 2002, 554).

Another prominent alternative to FE are regression models with clustered standard errors (Hillmert et al., 2017, 269). These models relax the assumption of independent residuals in linear models—i.e., heteroscedasticity—and allow for correlation of observation within clusters such as survey sample points or shared spatial locations and neighborhoods. Clustered standard errors provide adjustment for overestimated standard errors without requiring overly burdensome modeling techniques. Abadie, Athey, Imbens, and Wooldridge (2017) recommended to use clustered standard errors on every data that was part of a clustered sampling approach, such as the two survey data sets used in this book (see Chapter 2.3).

While ML models, FE models, and models with clustered standard errors provide alternatives to simple OLS regression models, issues of endogeneity may remain. None of these modeling techniques provides a solution to this issue. In cases in which geospatial data attributes stem from aggregated person-level data, endogeneity might confound estimates. The following section presents alternatives to ML, FE and clustered standard error models: *Spatial Econometric Models*. As such models are less common in social science survey research, the section provides more detailed information than the remarks above. Spatial Econometric Models represent an alternative in cases in which endogeneity states a problem.

5.3.2 Spatial Econometric Models

Alternatives to commonly used regression models for the analysis of spatially clustered data are *Spatial Econometric Models* (LeSage & Pace, 2009). Their emphasis is on providing the most efficient and unbiased estimators for spatially dependent and endogenous data. However, their specification is different in comparison to other regression models because, instead of using cluster information of observations' locations, they use the actual locations in the form of geo-coordinates. The following explains the idea of using geo-coordinates in regression models in more detail.

The general concept of Spatial Econometric Models is to connect observations in the data based on their locations. Although this is not necessarily different from ML models, the connection occurs on a case-by-case basis (Elhorst, 2014, 7). Thus, in-

stead of concentrating on the geographic clusters in which the observation i and j are located, as done in ML models, researchers connect these two observations through their geographic relation to each other. This geographic relation may comprise, for example, geodesic distances or adjacencies of neighborhoods (Neumayer & Plümper, 2016, 179). By using these connections, researchers can create lagged variables of observation i's outcome, covariates or even the error term that affects observation j's outcome. Spatial Econometric Models provide similar modeling opportunities as panel models, but instead of time, they use space as the basis for the lagged variables.

Accordingly, Spatial Econometric Models mimic spatial dependency on an observational level. In contrast to ML models, users of this approach do not compare observations on a group-by-group basis but an observation-by-observation basis. Moreover, while ML models require specific group and within-group observation numbers, Spatial Econometric Models are more flexible in this regard. They work on small-scale geographic locations such as point geo-coordinates of addresses or small neighborhoods.

Moreover, Spatial Econometric Models allow to distinct within-neighborhood and between-neighborhood effects (Dietz, 2002, 541). Within-neighborhood effects are comparable to ML modeling in which attributes on the group level are of interest, whereas between-neighborhood effects also take into account considerations of spatial spillovers or mutual influences of neighborhoods on each other. One of the advantages of Spatial Econometric Models, therefore, is the adjustment of spatial clustering of observations and elimination of estimates' biases. Also, researchers can analyze spatial diffusion processes between observations. For example, they can investigate how the outcome y_i of observation i affects the outcome y_j of observation j. Spatial Econometric Models provide a rich and flexible framework to incorporate an extensive range of different assumptions about spatial processes between people and neighborhoods.

Despite these advantages, specifying connections between observations in data requires to work on their actual conceptualization. Depending on which specification researchers choose, results of analyses may differ (Neumayer & Plümper, 2016). Although LeSage and Pace (2014) challenged this notion in recent years, it remains crucial to inspect substantial consequences of different concepts. For this reason, the following section introduces the more formal idea of building spatial connections between observations in data. It concerns the techniques of how to establish spatial connections, what these connections substantially mean and how the connections can be used in analyses.

The Connectivity Matrix W and Spatial Lags

Before researchers can use the flexible framework of Spatial Econometric Models, they have to connect the observations in their data. These connections are the basis for modeling spatial diffusion or spillover effects between observations (Elhorst, 2014, 6). In principle, these spatial connections are comparable to connections between observations in general network data. Each observation i connects to each observation j via a specific function, and while non-spatial interactions can be used as a basis for such a connection function (Neumayer & Plümper, 2016, 179), most of the time, this is done by using information about geographic locations.[27] Therefore, whenever researchers connect two observations, they assume some common exposure or social interaction between observations which is based on geographic proximity.

This connection of observations, based on geography, is represented in a rectangular $n \times n$ matrix, called the connectivity matrix W (Neumayer & Plümper, 2016). Each cell of the upper and the lower triangle of W depicts the corresponding connection's value of observation i vs. j and j vs. i, respectively—in Figure 5.8, this value is either 0 or 1. In contrast, the diagonal comprises a vector of zeros as each observation cannot be connected to itself. A rectangular connectivity matrix W for a dataset of observations contains, in principle, all essential information for the spatial analysis.

	i	.	.	.	n
i	0	1	0	0	1
.	1	0	0	1	0
.	0	0	0	1	1
.	0	1	1	0	1
n	1	0	1	1	0

Figure 5.8: Example of a Connectivity Matrix W with Binary Connections (0 = Absent; 1 = Existing)

The matrix in Figure 5.8 is an example of a binary connectivity matrix W for 5 observations. Either connections between observations exist, or they do not, resulting in a value of 1 for connections or 0 for non-existing ones (Elhorst, 2014, 10). Accord-

27 Moreover, Spatial Econometric Models were developed within the scope of the geographic data analysis. In such a setting, observations in data are geographic regions, which is why researchers often refer to observations in data as neighbors. In survey research, this is different: observations are individual people and not necessarily neighbors in the sense of geographic regions.

ingly, comparable to the diagonal of zeros for each observation's own cell combination, non-existing connections between i and j take on the same value, namely 0. In more complicated cases, as shown below, W can be more elaborated resulting in values of a larger range. Generally, W contains values for connections and non-existing connections, either in a binary format or in a more elaborated form.

The actual connections in W can stem from different approaches that also depend on the geometric properties of the observations. Connections can, for example, derive from contiguous neighbors of observations. If observation i's neighborhood A shares a mutual boundary with observation j's neighborhood B, establishing a connection may be appropriate. Alternatively, if these two neighborhoods do not share a mutual boundary but are within short proximity, establishing a connection may be appropriate as well. Connection modeling offers an extensive set of different methods that all rely on the earlier presented advantages of a projection of the observations into a joined coordinate space.

Meanwhile, one of the most prominent approaches is to use proximity as a direct indicator of observations' spatial connections (Hillmert et al., 2017, 270). This approach no longer results in binary connection indicators of zeros and ones but in geodesic distances between the two observations i and j. Whereas the resulting cell values aim to express the intensity of a connection, the value of geodesic distances increases with the distance between observation sites. The idea of geodesic distances is to operationalize the intensity of connections. For this reason, what researchers often do is to transform the value of geodesic distances through their inverse:

$$Distance_{inverse} = 1/Distance \tag{5.1}$$

With an inverse transformation, the higher the value of the geodesic distance, the lower is the inverse value. The values of connections in W are higher the closer the observations i and j are to each other. Thus, the inverse transformation converts connection values into the form intended by building W in the first place.

Figure 5.9 displays the effect of an inverse transformation on W. It first shows a connectivity matrix with geodesic distances between observations (Figure 5.9 (a)): the further away two observations are, the higher is their shared value. After applying the inverse transformation, the pattern changes (Figure 5.9 (b)): the further away two observations are, the lower is their shared value. In other words: the closer they are, the stronger is their connection.

The inverse transformation can be expressed in a distance decay function of influences (Hillmert et al., 2017, 270), displayed in Figure 5.10. The decay of the inverse transformation increases comparatively fast. Moreover, such functions are not lim-

	i	.	.	.	n
i	0	43	122	243	99
.	43	0	376	22	841
.	122	376	0	30	7
.	243	22	30	0	47
n	99	841	7	47	0

(a) Distance, e.g., in meters

	i	.	.	.	n
i	0	.023	.008	.004	.010
.	.023	0	.003	.04	.001
.	.008	.003	0	.034	.143
.	.004	.04	.034	0	.021
n	.010	.001	.143	.021	0

(b) 1 / Distance

Figure 5.9: Example of a Connectivity Matrix W with Distance and Inverse Distance Based Connections

ited to simple inverse functions. Researchers can use other functions such as inverse quadratic, gaussian or inverse in 10 meters steps transformation. Choosing one of these functions may have significant consequences for the intensity of connections: they can accelerate (inverse quadratic) and decelerate the decay (gaussian, inverse in 10 meters) or they influence the decays' trajectories shapes (inverse in 10 meters).

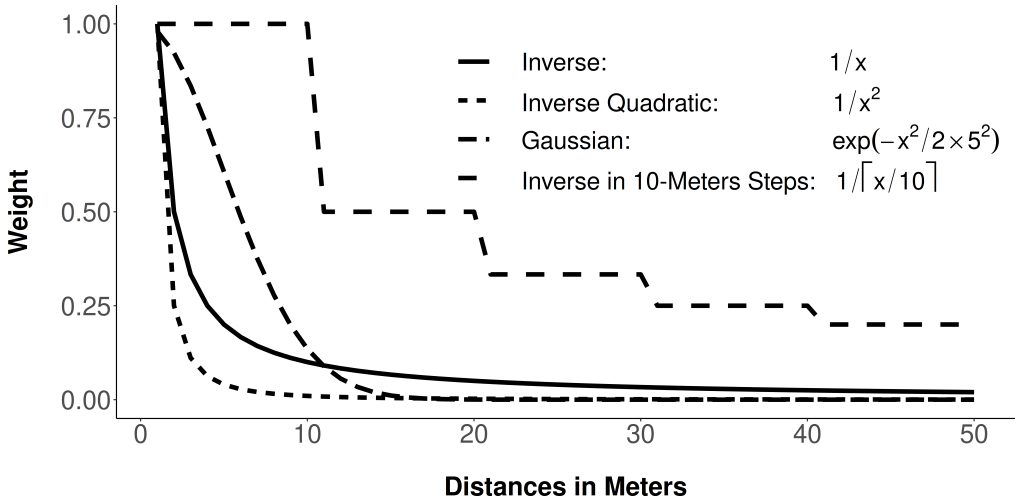


Figure 5.10: Different Decay Functions Depending on Distances and Their Impact on Weights

There is no strict rule of thumb which of the functions researchers ought to use. Instead, Neumayer and Plümper (2016) recommend consulting the theory to justify the choice of a specific modeling approach. The empirical application in Chapter 7, for example, uses an inverse in 10 meters transformation of the geodesic distances between survey respondents' geo-coordinates. Because of the random sampling of surveys in

municipalities, clustering of respondents within single settlements is not common. Figure 5.10 shows that already after a few meters distance between respondents the weight is significantly decreasing. Thus, using transformations, such as the simple inverse, results in only a few influential weights, but the dependent variable of the analyses, xenophobia, is hypothesized to cluster within larger neighborhoods. For this reason, the inverse in 10 meters steps is chosen as the decay accelerates slower than the other transformations.

Regardless of any transformation, connectivity matrices remain to be rectangular matrices whose number of columns and rows depends on the number of observations in the data. Spatial Econometric Models for regression analysis, on the other hand, are still regular regression models using a standard design matrix D with rows depicting the observations 1 to n and columns presenting the variables 1 to k. The question, thus, remains how the information from the $n \times n$ matrix of W can be included in the $n \times k$ matrix of D. For this purpose, W gets normalized so that it results in a column vector of length n (Elhorst, 2014, 12f). Each cell of this normalized vector of W then renders a summary measure of influence for the observation j on observation i. In order to include connectivity matrices W in Spatial Econometric Regression Models, their information is reduced in the form of summary statistics.

Spatial Econometric Models use these summary statistics of connections in combination with observations' measures of interest. Depending on the mechanism of spatial influence—spillover, diffusion or common exposure—the reduced form of W is multiplied with the outcome of i's neighbors, their covariates or their error term. Each one of these combinations results in single coefficients in a regression, indicating the overall neighbors' influence on the outcome, covariates or error term of observation i.

Before using such modeling approaches, researcher have to decide between several options to normalize W. These options differ with regards to their influence on the actual weight each neighbor j has on i. For example, some normalizations balance each connection value w_{ij} against the number of all other connections, some of them relate them to extreme connection values such as maxima values. One of the frequently used methods, row normalization (Elhorst, 2014, 12), belongs to the former class:

$$W_{row} = \sum_j \frac{w_{ij}}{\sum\limits_j w_{ij}}, \tag{5.2}$$

where i = 1, 2, ..., n and j = 1, 2, ..., n. Thus, w_{ij} depicts the connectivity indicator derived from each cell combination exemplified in the matrices of Figure 5.8 and 5.9 above. By dividing each combination through the sum of all combinations, a standard-

ized measure of the relative weight of each ij combination for each row, summing up to 1, is retrieved.

As described above, in Spatial Econometric Regression Models the normalized weight is multiplied with the parameter of interest, e.g., the outcome of i's neighbor y_j. If researchers have an interest in a lagged neighborhood outcome effect, they multiply the relative weight w_{ij} with y_j. To estimate an overall coefficient for this lagged effect, they subsequently sum up all multiplications across j. Consequently, this summation results in a model of lagged neighbor outcomes as follows:

$$y_i = \rho \sum_j \frac{w_{ij}}{\sum_j w_{ij}} y_j + \beta X_i + \epsilon_i. \tag{5.3}$$

Such a model is comparable to a standard linear regression model with a $D = \beta X$ design matrix and a disturbance term ϵ. Note, however, the inclusion of the outcome y_j of the observation j which is multiplied by the row standardized weight. This procedure estimates the parameter ρ, an overall measure of the spatial influence of neighbors' outcomes on the outcome of observation i. Spatial Econometric Models generally differ from common regression models by estimating additional parameters for spatial influence with the help of the connectivity matrix W.

The following section presents this specific model in more detail. This model depicts a commonly used Spatial Econometric Regression Model, which is also used in the empirical application of Chapter 7. It belongs to the family of Spatial Lag Models in which a connectivity matrix W determines the spatial relationship among observations. Thus, this elaboration is not exhaustive, but it already covers a large part of different applications in research.

Spatial Lag Y Models

The previous section already introduced Spatial Lag Y Models that conceptualize spatial relationships across observations with regards to their outcome (Elhorst, 2014, 7).[28] They establish the spatial relationships for the observation i by multiplying the outcomes of each observation j with the connectivity matrix W. The multiplications are averaged with the result that the parameter ρ comprises an overall measure of the influence of i's neighbors' outcomes on i's outcome.

The general idea of Spatial Lag Y Models is that changes in the outcome of one ob-

28 In the literature, a large number of different names for the same or similar Spatial Econometric Models exists. For example, some authors call Spatial Lag Y Models simply Spatial Lag Models or Spatial Autoregressive Models (Elhorst, 2014, 5).

servation j also lead to changes in the outcome of another observation i. Comparable to panel data models in which an earlier measured outcome of one observation affects its outcome in the future through a lagged coefficient, a lagged coefficient determines the relationship here as well. The difference is that the lag does not occur within individuals but between them through their spatial relation. Furthermore, whereas time moves in one direction—from past to future—, spatial effects occur between the observation j on i and i on j, thus vice versa. Spatial Lag Y Models create feedback effects between observations in space.

Such feedback models may apply in different theoretically meaningful situations. They are meaningful for theories which assume crossover effects of social interactions that self-enforce people's responses, attitudes, or characteristics. For example, people might talk about political issues with their neighbors. According to *Social Contagion Theory* (Christakis & Fowler, 2013; Steenbeek & van Geert, 2007), they might end up with similar opinions on specific topics because of their mutual influence on each other. If one person i changes her or his opinion, a change of opinion of her or his neighbor j might become probable as well because of social interaction between i and j. In turn, the change of opinion of neighbor j can affect observation i's opinion again with the result of a feedback effect. Crossover effects between neighbors exemplify the spatial dimension of social interactions and feedback effects with regards to people's outcomes.

Above, Equation 5.3 already introduced the notation for Spatial Lag Y Models which shows how these models incorporate the spatial feedback effects. They estimate an additional parameter called ρ which is the overall spatial influence of neighbors' outcomes on y_i. In contrast to panel models where the outcomes of previous time points affect the outcome of subsequent time point, these feedback effects occur through space and in two directions.

Spatial Lag Y Models, therefore, are endogenous by design because elements of the left-hand side of the formula turn up as elements of the right-hand side of the formula as well. Using estimators such as OLS would fail to compensate for this endogeneity. Instead, other estimators such as Two-Stage Least Squares (Elhorst, 2014, 17) or Line Searching for spatial regression (Bivand et al., 2008, 306) are more appropriate to apply because they specifically address endogeneity in their algorithm. Most of the statistical software for spatial regression models implemented these estimators in their packages, as such this section refrains from differentiating between them. Most importantly, Spatial Lag Y Models and software implementations offer methods to compensate for endogeneity in estimations.

5.3.3 Choice Between Models

After presenting commonly used regression models and Spatial Econometric Models, one question remains: what specific type of these models should researchers choose for their analysis? The first answer is that the theory should justify the modeling of spatial relationships of observations in data (Neumayer & Plümper, 2016, 191), given that there is spatial dependence at all (Elhorst, 2014, 7). Consulting the theory is the first step in assessing the sources of spatially dependent data.

The theory is of particular importance in the case of georeferenced survey data and spatial linking. They imply linking of two data sources that are often not part of a shared data generation process. While it is true that, for example, soil sealing is a consequence of human behavior, its specific manifestation is not related to questions asked in a social science survey. Its attributes were gathered using satellite image recognition techniques, and not through survey questions. For this reason, it is questionable whether spatial dependence between survey respondents with regards to such attributes exists and whether it has to be controlled.

This confinement, however, changes when survey attributes are spatially clustered, particularly among dependent variables. For instance, the empirical application in Chapter 7 uses, amongst others, a Spatial Lag Y Model for the dependent variable of political attitudes. Its implementation is appropriate because the theory assumes social interactions between people with regards to political attitudes. Also, statistical tests reveal spatial dependence between the survey respondents. In all other applications, the dependent variables were either survey measures without social interactions between respondents or spatial attributes that were not spatially related to other respondents.

Besides these essential theoretical considerations, there are also statistical tests which can help to detect spatial dependence (Bivand et al., 2008, 258ff) and to assess different sources of this spatial dependence (Elhorst, 2014, 5). Conceptually, spatial lag models are nested in each other and in the general OLS model, whereby in OLS models all potential spatial lags are assumed as being zero. OLS, therefore, depicts the most constrained model (Elhorst, 2014, 9). Researchers can gradually weaken these constraints and test whether a specific model fits better with the data.

One remaining methodological issue with Spatial Econometric Models is that the connectivity matrix W assumes full information on geometries (Gibbons & Overman, 2012). Comparable to network analysis, missing links between neighbors would distort the relationships between other neighbors, depending on the construction of the matrix. For georeferenced survey data, the circumstances are rather severe: these data consist of a random sample of people from specific areas, but people who were

not part of the random sample are not interviewed. Connectivity matrices based on, e.g., 4 people who live close to each other would render them as neighbors, but the actual neighbors who were left out in the sampling process are skipped. However, using these techniques, at least implicitly, assumes that there are actual relationships between these neighbors.

In some cases, using models which allow clustering of survey respondents in geographic space may pose an adequate alternative to Spatial Econometric Models. Förster (2018) used information about 1 km^2 census grid cells to create spatial clusters between survey respondents of the German Longitudinal Election Study (GLES). The respondents were assigned to the grid cell in which they lived. This information was used to estimate ML models which included four different geographic clusters, reaching from the small scale 1 km^2 grid cells up to administrative districts. This example shows that, even for spatial clustering on different geographic levels, commonly used regression models may provide a valid modeling approach.

To sum up, researchers should critically examine all available options. These options depend on the research question, on associated theories about spatial processes and, not least, on the actual data. Not all spatial attributes, gathered through spatial linking methods, require the application of elaborated Spatial Econometric Models. By and large, spatial attributes are additional variables in survey data and do not necessarily impose any needs to change conventional modeling approaches. Only one of the following empirical applications uses an analysis comprising an actual Spatial Econometric Model. In all other applications, controlling spatial dependence between respondents was possible to achieve with commonly used modeling techniques, such as clustered standard errors or FE models.

6 Application I: Road Traffic Noise, Marriage, and Health

The present chapter portrays the first empirical application of georeferenced survey data in this book. As the other following empirical applications in Chapter 7 and 8, it connects the information from previous chapters by deploying spatial linking methods to relevant social science research questions. Each empirical application also differs from the others by being associated with different social science disciplines and by using various methods. The purpose of the applications is to exhibit the variety and the manifold possibilities of georeferenced survey data.

The empirical application at hand made use of the GGSS (Chapter 2.3.1) spatially linked to geospatial data on road traffic noise (Chapter 2.4.2). To apply spatial linking to its full potential, two methods were introduced: spatial linking by location (Chapter 5.2.1) and spatial linking with geodesic distances (Chapter 5.2.4). In this sense, this empirical application exemplifies potential snares of using georeferenced survey data that can be navigated by knowing the involved data as well as associated methods.

6.1 Research Question

Are married people less likely to report health problems when they are exposed to the deleterious effects of road traffic noise stressors? This question refers to a long debate in the social sciences involving the discussion of how marriage influences health. Although societies have changed relating to the quantitative and qualitative differences between married and unmarried couples (Liu & Reczek, 2012), married persons are still found to be at better health than their unmarried counterparts. Thus far, recent findings showed that married people, for instance, live longer (Drefahl, 2012) and married men, in particular, are at lower risk to develop myocardial infarcts (Kilpi, Konttinen, Silventoinen, & Martikainen, 2015). While some authors argued that this health effect of marriage is due to, among others, a reduction of unhealthy practices within marriages, specifically of men (Carr & Springer, 2010, 750), evidence in support of a direct effect of marriage on health exists (Robles, Slatcher, Trombello, & McGinn, 2014, 141).

What frequently was suggested to be one prominent cause of this direct health effect of marriage is stress buffering (Thoits, 2010). Researchers demonstrated that not only dyadic coping of married couples is more effective (Bodenmann, 1997), but also that direct bodily pathways such as stress hormones or neural responses of married people are different to their unmarried counterparts (Coan, Schaefer, & Davidson,

2006; Maestripieri, M. Baran, Sapienza, & Zingales, 2010). These findings imply that married couples share resources and interactions other people in different living arrangements do not have access to, which, in consequence, directly advance physical and mental health through stress buffering (Meyler, Stimpson, & Peek, 2007, 2298).

Yet, of all stressors intruding both married and unmarried people's lives and homes (Serido, Almeida, & Wethington, 2004), researchers know little about the differences of stress buffering between married and unmarried people towards the deleterious effects of external stressors such as the exposure to environmental noise. The latter was identified as one considerable risk factor for developing mental and physical illness (Passchier-Vermeer & Passchier, 2000). Also, because the vulnerability to environmental noise was found to differ individually (Stansfeld & Shipley, 2015) and also between different demographic groups (Brink, 2011; Kohlhuber, Mielck, Weiland, & Bolte, 2006), it is surprising that no systematic empirical evidence exists exploring health differences between married and unmarried people who are exposed to environmental noise. Given that marriage advances buffering of stressors, it could be argued that, indeed, differences of the deleterious effects of environmental noise between married and unmarried people may be prevalent.

This empirical application thus compares health outcomes among married and unmarried people who are exposed to road traffic noise. It conceptualizes road traffic noise stressors as part of the onset of stress, corresponding to a general stress process model that distinguishes stressors, intervening factors and health manifestations (Aneshensel, 2015). According to this model, marriage intervenes and buffers the general adverse effects of stressors, such as road traffic noise, on mental and physical health manifestations.

6.1.1 Marriage as Stress Buffer

Building upon the literature on social ties and health (Berkman et al., 2000; Thoits, 2011) as well as on the work on social influences of stress processing (Aneshensel, 2015; Pearlin, Schieman, Fazio, & Meersman, 2005), and dyadic coping (Bodenmann, 1997), the following section describes a combined general model for the association between road traffic noise, health, and marriage.

The model assumes that social relationships influence health through different mechanisms, such as social influence, social control, and social support (Thoits, 2011). These mechanisms either directly approach the harmful effects of stressors, or they help by preventing and buffering the emergence of the deleterious effects of stressors. By providing active social support or by promoting personal resources

such as self-esteem and mastery, manifestations of the stressor in the form of, e.g., depressive symptoms are intercepted and mediated (Aneshensel, 2015, 167f). In consequence, individuals exposed to a potential stressor actively use available resources to cope with the situation, reassess its source and perform preventive actions, for instance, also by avoiding stressors (Thoits, 1994).

Another prominent mechanism through which social ties intervene in the stress process concerns people's perception of stressors as being stressful. Social ties advance personal resources such as the sense of control, but they also strengthen the feeling of belonging and companionship. These resources have been found to have a direct link to the processing of stress in the human brain (McEwen, 2008, 181). Because these resources do not only exist during episodes of stress, they also tend to be preventive with regards to the perception of future stressors. As a result, individuals who are exposed to a potential stressor and who can draw on resources provided by social ties perceive the stressor as less stressful in comparison to individuals who cannot draw on these resources (Maestripieri et al., 2010, 422).

Following these arguments, different interconnected explanations are elucidating why married couples, in particular, are generally better in coping with stressors and, in consequence, are healthier than unmarried people. For the research question of this empirical application two explanations are of high relevance:

First, married couples differ from unmarried couples with regards to the amount of available coping resources (Liu & Reczek, 2012). For example, by social integration into extended families, marriage increases, besides economic resources, the available access to psychological resources such as social support. As previous research has shown empirically, unmarried people receive less support from friends or relatives of their partner (Liu & Reczek, 2012, 796; Drefahl, 2012, 472).

Second, married couples report higher levels of relationship satisfaction than unmarried couples. Also, unmarried couples experience higher levels of conflicts within the relationship, are more insecure concerning the relationship status and are more likely to introduce severe mental strains into the relationship (Liu & Reczek, 2012). Therefore, since coping resources are affected by higher levels of conflict and strains, internal and external stressors cannot be adequately addressed by unmarried couples.

Indeed, various other explanations may apply. For example, it was hypothesized that healthy people select themselves into marriages. It was also found that non-healthy people are more likely to get divorced (Drefahl, 2012, 462f). Another explanation refers to differences in the health behavior of married individuals, such as reduced alcohol or tobacco consumption (Carr & Springer, 2010, 749). However, because this application studies the influence of road traffic noise and its perception,

in the later analysis, these factors are, whenever this is possible, included as control variable.

Overall, it is important to note that all these explanations and findings describe a general and gradual tendency of differences between married and unmarried people. Studies conducted before, e.g., revealed that relationship satisfaction is one of the most important predictors of effective dyadic coping; both for married and unmarried couples (Kouros & Cummings, 2010). This application, therefore, hypothesizes a tentative positive effect of being married when individuals are exposed to road traffic noise.

6.1.2 Road Traffic Noise, Health and Its Link to Marital Status

After describing the pathways through which marriage buffers the effects of stressors, the following establishes the link between marriage and its influence on road traffic noise stressors.

As stated earlier, environmental noise has a strong influence on people's health (Passchier-Vermeer & Passchier, 2000). Recent research corroborated links to hearing problems (Basner et al., 2014), cardiovascular diseases (Babisch et al., 2014) and even diabetes (Sørensen et al., 2013). Because road traffic noise is the most prevalent source of environmental noise, its effects are noticeable among a broad range of the population.

While the results regarding physical health and road traffic noise are considerable, the findings regarding mental illnesses such as depression or anxieties (Stansfeld et al., 2000) are ambivalent. The same holds for subjective measures of health; many studies hardly find any effects at all. Roswall et al. (2015), for instance, concluded that lifestyle factors thoroughly explain their observed but weak associations between road traffic noise and mental health. Hardoy et al. (2005), on the other hand, showed earlier that under consideration of subclinical symptoms and even though their study was on air traffic noise, indeed, strong associations can be found.

What a large set of explanations identify as a critical factor for understanding these ambivalent outcomes is environmental noise perception. For example, in a recent study, Heritier et al. (2014) demonstrated that environmental noise perception mediates the effects of road traffic noise on health. Based on the assumptions of stress research it can be assumed that road traffic noise only emerges as a stressor when people perceive it as an actual threat.

This empirical application hypothesizes that marriage influences individuals' road traffic noise perception. Referring to the concept of direct and preventive buffering,

it further assumes that this includes both, knowingly and unknowingly, road traffic noise perception. Hence, it also expects that the effect of road traffic noise on its perception differs between married and unmarried people.

6.2 Data and Methods

6.2.1 Geospatial Data Measures

Testing the theoretical considerations outlined above requires combined survey data and road traffic noise data. For the survey respondents, measures on health and marital status, as well as noise measurement values at their dwelling, are necessary. While the first set of characteristics derives from survey data that are described after this section, the latter can be obtained from the spatial linking procedures. Therefore, the following section begins with the efforts of spatial linking for the two data sources.

Data Considerations

The road traffic noise measurements stem from the data described in Chapter 2.4.2 which were collected in correspondence with the Environmental Noise Directive (2002/49/EC) of the European Union (EU) (European Parliament & European Council, 2002). Also, Chapter 2.4.2 already demonstrated that the intention of data collection was conceptualized according to a central EU directive, but the implementation in German law has led to distributed responsibilities regarding the actual data collection. Some federal states collected data for smaller municipalities, but most of the larger municipalities had to collect the data on their own.[29] As a consequence of this implementation of the EU directive, no central publisher gives access to the whole set of the road traffic noise data in Germany.

Furthermore, a side-effect is the fact that the data which are accessible are not harmonized. They exist in different data formats, the noise measurements relate to different geometries, and attribute categories vary. Where possible, users of the data may have to harmonize them; however, that is not always practicable (Schweers et al., 2016). Although preparing these data for spatial linking requires high efforts in harmonization, the data are still useful for a majority of survey respondents and the intended spatial linking.

Even so, in the course of interpreting the results of the analyses, discrepancies be-

29 https://www.gesetze-im-internet.de/bimschg/BJNR007210974.html, § 74

tween the input data have to be considered. These discrepancies can lead to differences in effect sizes as well as variances and are a potential source of error and unobserved heterogeneity. Although noise was measured the same way at least in each German municipality, heterogeneity can still occur between them. Therefore, it is crucial to apply statistical models that account for this heterogeneity. One way to address these issues is, for example, using cluster-robust standard errors in regression models (see Chapter 5.3.1).

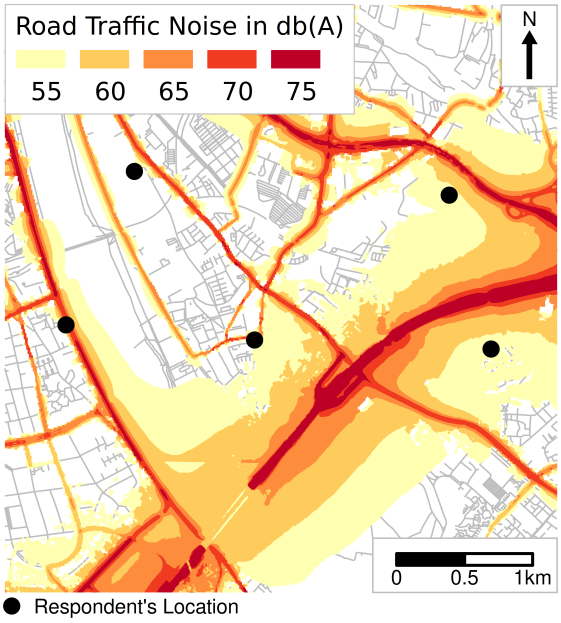
Before the results of the analyses are presented below, the next two sections describe the procedure of spatial linking. Because there are different options of spatial linking, it depends on the data which method is appropriate for a specific use case (see Chapter 5.2). For the introduced research question, spatial linking by location and geodesic distances qualify.

Spatial Linking by Location of Road Traffic Noise

This book's road traffic noise data depict detailed geographic information. While their noise values are available as categorized dB(A) ranges, the geographic extent is even more sensitive to change. Point A is probable to have a different value than point B which is located 10 meters to its south. On point A, the road traffic noise measurement may depict 65 dB(A), whereas point B only reaches 50 dB(A). Consequently, these detailed data provide high flexibility for spatial linking.

The simplest way to combine georeferenced survey data and road traffic noise data is spatial linking by location (see Chapter 5.2.1). Because the road traffic noise measurements vary even within the same neighborhood, they can take on high and small dB(A) values within one small area. The reason is that different roads in one neighborhood are frequented distinctly, and also obstacles such as buildings or trees may diminish sound waves between people's locations. Using the road traffic noise values measured at the facade of a building may be the most viable approach for spatial linking.

Figure 6.1 visualizes the approach of spatial linking by location of road traffic noise data and the GGSS 2014. A map section of the city of Cologne displays a road traffic noise layer for measurements on the main streets of the city. In the middle of the streets, the measurements yield higher dB(A) values, whereas they are lower the further away they are from the center of the streets. The black points in the map represent fictional respondents' housing locations of the GGSS. The spatial linking by location procedure assigns the corresponding road traffic noise measurements to the respondents' location.



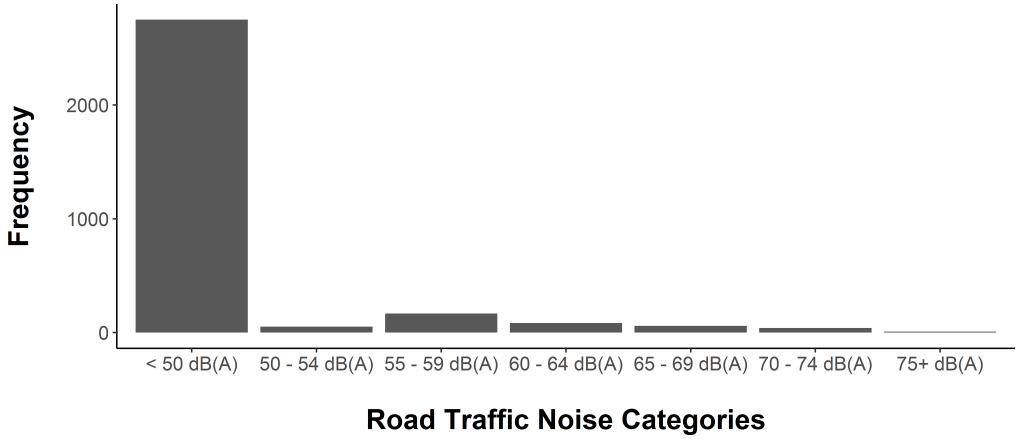
Data Sources: German Environmental Agency / EIONET Central Data Repository (2016) and OpenStreetMap / GEOFABRIK (2018)

Figure 6.1: Spatial Linking by Location of Road Traffic Noise Data and the GGSS 2014

Though, not all respondents receive an actual value for the measurement of road traffic noise. Plenty of respondents within the area do not live close to a main road, but only these roads were mapped in correspondence with the EU directive. Furthermore, measurements often stop at the facade of the buildings, however, the respondents' geo-coordinates sit in the center of buildings. Accordingly, for all respondents that fulfilled at least one of these conditions, no measurement of road traffic noise is found.

While Figure 6.1 displays fictional survey respondents' locations, the distribution of the results of spatial linking by location for the real data corroborates the suspicion of missing measurements. Solely a small proportion of GGSS 2014 respondents receive a noise value above 50 dB(A) at all, 88.33 percent of the respondents have a measurement between 0 and 49 dB(A). The distribution is skewed, and the scoring of dB(A) range categories is low, as Figure 6.2 shows. In consequence, applying statistical procedures such as regression models may lack statistical power.

As the results below show, analyses still yield valuable results. At the same time, it is essential to assess them according to their robustness; and spatial linking provides a flexible toolkit to do so. Thus, the following section presents statistical adjustments using another procedure of spatial linking for the above-presented research question.



Data Source: Georeferenced German General Social Survey 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Figure 6.2: Distribution of Road Traffic Noise Attributes Retrieved from Spatial Linking by Location (N = 3,163)

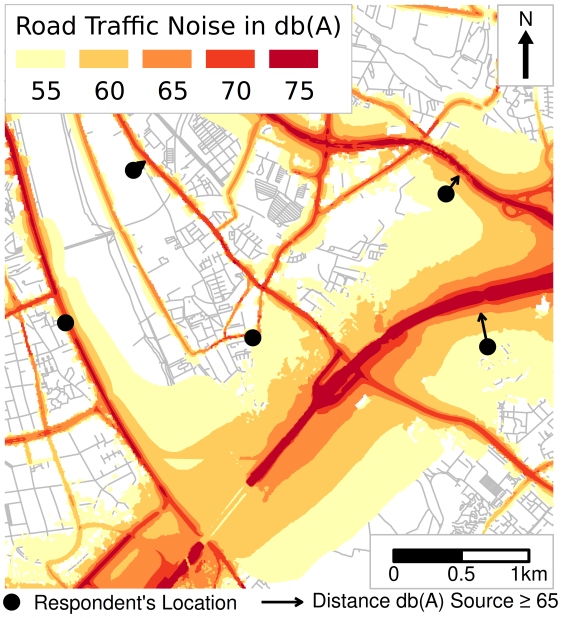
Geodesic Distances to the Next Road Traffic Noise Source

To navigate the issues of spatial linking by location, the use of geodesic distances is helpful. The road traffic noise data are unambiguously censored—first, they do not contain values below 50 dB(A), and second, they contain just a few measurements for neighborhoods with low traffic densities. Building on the hypothesis that the closer persons live to a specific noise source, the more they are exposed to their deleterious effect, geodesic distances can serve as a cue to the general noise exposure in a specific neighborhood. Moreover, auxiliary roads may be more frequented the nearer they are to main roads, so that geodesic distances could also capture respondents whose locations were not measured, but, at least, live within neighborhoods with potential noise exposure. Accordingly, geodesic distances can serve as *Instrumental Variable* (IV) (Angrist, Imbens, & Rubin, 1996) that increases the likelihood of exposure for respondents.

Another issue that can be navigated by geodesic distance calculations is the fact that road traffic noise measurements often stop at buildings' facades. Geo-coordinates which depict centroid coordinates of buildings' polygon geometries often yield zero values because of missing noise measurements. At the same time, it is unrealistic that people within such buildings are not exposed to any noise from roads. Accordingly, these geo-coordinates lead to measurement errors which may also be approached by instrumenting the original variable (Wooldridge, 2012, 532ff).

Comparable to Figure 5.7, Figure 6.3 displays the procedure of spatial linking, in

this particular case, with geodesic distances.[30] In this specific application, these geodesic distances are distances to the next noise source producing at least 65 dB(A), a level of noise which, according to the German Environmental Agency, is identified as a threshold to develop severe noise-related health issues.[31] In this sense, even if respondents do not live within a small radius of a noise source, geodesic distances below a certain value may signal a potential likelihood of exposure.



Data Sources: German Environmental Agency / EIONET Central Data Repository (2016) and OpenStreetMap / GEOFABRIK (2018)

Figure 6.3: Spatial Linking with Geodesic Distances of Road Traffic Noise Data and the GGSS 2014

The IV approach comprises two steps. In the first step, the original road traffic noise variable is regressed on the geodesic distances and all other covariates. In the second step, the predicted values of road traffic noise are extracted and used as a measure of road traffic noise in the actual analysis. This procedure discards all variance from the original road traffic noise variable which may pose a measurement error. As described above, these errors stem from missing data on auxiliary roads or measurements which stop at the facade of buldings. By using this approach, road traffic noise in the regressions should yield more accurate results.

30 The distribution of geodesic distances is illustrated in Figure A.1 in the appendix.
31 https://www.umweltbundesamt.de/en/indicator-population-exposure-to-traffic-noise

The applied IV regression approach can also be written as a set of structural equations for each dependent variable in the following analysis:

$$Y_i = \beta_0 + \beta_1 T_i + \sum \beta_i X_i + \epsilon_i \tag{6.1}$$

and

$$T_i = \alpha_0 + \alpha_1 Z1_i + \sum \alpha_i X_i + \upsilon_i, \tag{6.2}$$

where Y_i depicts the outcomes of the health measures or noise annoyance (see below), β_0 their estimated mean value, T_i the treatment of the road traffic noise measure, $\sum X_i$ a vector of control variables and ϵ_i the disturbance term. T_i is then regressed on the instrumental variable $Z1_i$, the geodesic distances, and on the covariates $\sum X_i$. This application uses the *Two-Stage Least Squares* (2SLS) approach (Wooldridge, 2012, 512ff) in which first the regression 6.2 is estimated and the predicted values are stored as a new measure of road traffic noise for the actual regression in 6.1. The result is a measure that no longer suffers from measurement errors and displays a more realistic picture of road traffic noise exposure for each survey respondent.

6.2.2 Survey Measures

Physical and Mental Health

Among others, the GGSS 2014 contains the short form version (SF-12) of the Health-Related Quality of Life (HRQL) questionnaire as a measure of physical and mental health. This questionnaire includes eight sub-scales on general health, physical functioning and physical roles, bodily pain as well as vitality, social functioning, emotional roles, and general mental health. Usually, these subscales are summarized into two scores—the Physical Component Score (PCS), and the Mental Component Score (MCS)—by using Principal Component Analysis (PCA). A research group of the SOEP proposed an algorithm to compare each new collection of this questionnaire with a norm-sample of the SOEP (Nübling, Andersen, & Mühlbacher, 2006). Accordingly, the HRQL represents a standardized measure of health for an extensive range of different applications in social science survey research.

However, in this application, the HRQL is not used in this manner. Instead, the z-standardized subscales[32] are directly entered in the analysis to find a confirmatory fac-

32 The corresponding survey questions in the GGSS 2014 can be found in Table A.1 of the Appendix.

tor solution. Moreover, while the common PCA approach assumes that the two component scores do not correlate with each other, in this application they do—a much more realistic assumption about the link between physical and mental health (Timmermans & Haas, 2008, 660).

Marital Status

One of the main independent variables of this application is a distinction between married and unmarried people. The GGSS 2014 contains legal definitions for each marital status group in Germany. This information was used to build a dichotomous indicator of the different groups with 1 standing for married people and 0 for unmarried people.[33]

Noise Annoyance

Noise annoyance has been identified as a relevant and crucial mediator between actual road traffic noise exposure and health (Heritier et al., 2014) and, at the same time, is the other main independent variable of this application. Noise annoyance was covered in the GGSS by two questions: "How strongly are you disturbed or annoyed through noise in your neighborhood during the day?" and "And how strongly are you disturbed or annoyed through noise in your neighborhood during the night?" (1 = "very strongly", 2 = "strongly", 3 = "fair", 4 = "little", and 5 = "not at all"). This scale was reversed so that smaller values mean less noise annoyance and larger values stand for more noise annoyance. In a later used latent variable approach, the latent variable depicts the general noise annoyance of survey respondents.

Sociodemographics

Past evidence regarding the determinants of people's health as well as potential confounders of sociodemographics and environmental noise motivates the choice of sociodemographic control variables. As age is known to be strongly related to health, the year of age is used as a control variable. Gender differences were also subject of a large body of research, therefore, in the analysis, a binary indicator of gender was included

33 People who live in a civil union were also assigned to the group of married people. The argument is that entering civil unions involves similar commitments as ordinary marriages. Meanwhile, this affects only four respondents of the whole sample; thus, even if a structural difference between civil unions and marriages existed, the bias would be negligible.

(0 = male and 1 = female). Education was recognized to be another important dimension of a social gradient of people's health. Hence, four levels of education based on the International Standard Classification of Education (ISCED) are used, ranging from low, medium, advanced up to high education. The income, operationalized as the logged income in EUR, presents a further important factor. Finally, as an indicator of health behavior that determines physical and mental health tremendously, a binary indicator for smoking habits is included in the analysis.

Municipality Size

Road traffic noise also depicts a phenomenon of urbanization. Thus, the analysis includes the size of the municipality each respondent is living in as a control variable (1 = "up to 1,999 inhabitants", 2 = "2,000–4,999 inhabitants", 3 = "5,000–19,999 inhabitants", 4 = "20,000–49,999 inhabitants", 5 = "50,000–99,999", 6 = "100,000–499.999 inhabitants", and 7 = "500,000 and more inhabitants"). In the analysis, as the effects of specific municipality sizes is not of interest, this variable is entered as continious indicator.

An overview of all variables used in the analysis is shown in Table 6.1. It describes the dependent variables PCS and MCS, the mediator noise annoyance, the central predictor of road traffic noise (before the IV approach adjustment), the moderator marital status and the control variables. The next section continues with the strategy of analyzing the data.

Table 6.1: Descriptive Statistics and Overview of all Variables of the Analysis (Listwise Deletion)

	Mean/%	SD	Minimum	Maximum
Physical Component Score (PCS)				
General Health	0	1	-2.54	1.55
Physical Functioning	.01	.99	-2.24	.80
Role Physical	0	1	-2.57	.87
Bodily Pain	0	1	-2.55	.87
Mental Component Score (MCS)				
Vitality	0	.99	-2.40	1.88
Social Functioning	-.01	1	-3.78	.57
Role Emotional	0	1	-3.71	.65
Mental Health	-.01	.99	-2.58	1.75
Noise Annoyance				
Annoyance During Day Time	1.78	1	1	5
Annoyance During Night Time	1.52	.84	1	5
Road Traffic Noise[a]				
< 50 db(A)	86.97			
50 - 54 db(A)	1.63			
55 - 59 db(A)	5.35			
60 - 64 db(A)	2.61			
65 - 69 db(A)	1.85			
70 - 74 db(A)	1.31			
75+ db(A)	.29			
Marital Status				
Married	.56	.50	0	1
Controls				
Age	49.26	17.40	18	91
Gender	.49	.50	0	1
Education	2.56	.96	1	4
Logged Income	6.74	1.77	0	11
Smoking	.29	.46	0	1
Municipality Size	2.57	1.14	1	4
Number of Observations	3138			

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); [a] original values before applying IV approach

6.3 Analysis Strategy: Structural Equation Modeling

The research question of this application implies complex relationships and interactions among different groups of people. Analyzing data according to such theories requires statistical models that can capture all relevant information comprehensively and unambiguously. A method which is adequate for this purpose is *Structural Equation Modeling* (SEM). SEM can include manifest (e.g., marital status) as well as latent variables (e.g., mental health), and can test relationships across multiple independent and dependent variables (Schumacker & Lomax, 2010, 180ff). To test the hypotheses of mediating and moderating factors in the theory of stress buffering the use of SEM is promising.

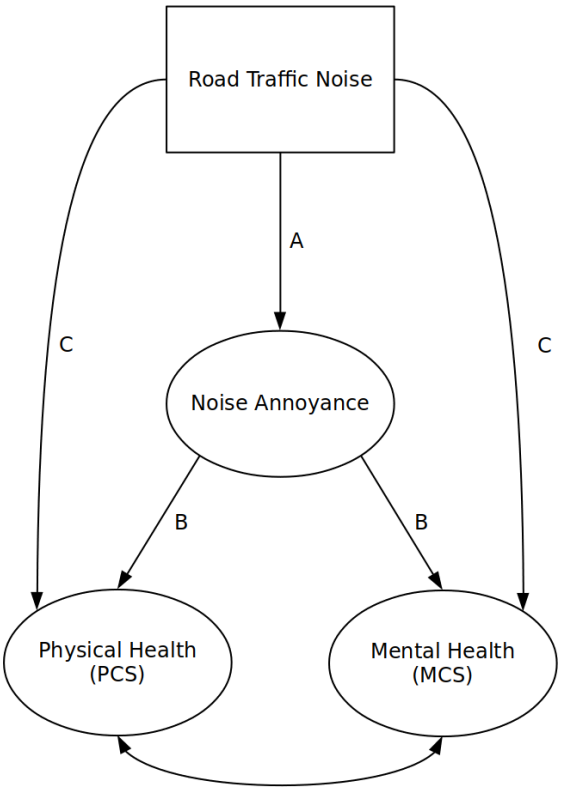


Figure 6.4: SEM of the Relationship Between Road Traffic Noise, Noise Annoyance, and Physical and Mental Health

For this purpose, the relationships between road traffic noise, noise annoyance, and physical and mental health are modeled as shown in Figure 6.4. Road traffic noise as independent and manifest variable affects the latent variables physical and mental health directly (paths C) and indirectly (path A × path B) through the mediating effect

of the latent variable noise annoyance on physical and mental health. Not included in this path diagram are the control variables that aim to take extra variance in the potential relationships into account. All in all, the estimated SEM depicts a classic mediation model that allows to statistically test direct and indirect paths between the variables of interest.

What is still missing from this model is the moderating effect of the marital status of respondents. This interaction is realized by estimating a multigroup SEM (Schumacker & Lomax, 2010, 250). Accordingly, the model from Figure 6.4 is estimated simultaneously for the two subsets of people in the data: married and unmarried people. An advantage of this approach is that separate paths in the model can explicitly be tested for their statistical difference. Mainly, with regards to the hypotheses of this application, differences among the paths C, A, and B are of interest. Moreover, it can be analyzed whether the indirect effects (path A $\times$ path B) or even the total effects of road traffic noise on each health dimension (path C + path A $\times$ path B) differ. The SEM approach, with its possibility of multigroup analysis and separate path constraints, represents a flexible method to test differences between married and unmarried people in the complex nexus between road traffic noise and health.

All SEM models are estimated using the *R* package *lavaan* written by Rosseel (2012). As estimator, *Robust Maximum Likelihood* (MLR) is chosen because it behaves reliably in case of non-normally distributed data and categorical data while, at the same time, coefficients are not overestimated (C.-H. Li, 2016). Moreover, the standard errors depict cluster-robust standard errors for sample points of the GGSS to adjust for the dependence of observations within sample points (Abadie et al., 2017).

6.4 Results

6.4.1 Model Fit

The following presents the model fit of the SEM which is necessary to assess its general reliability. It is also essential for determining how much the model fits the data because, in any SEM application, researchers aim to find evidence that modeling of the paths makes sense. Table 6.2 presents the results for this application's model. Generally, the model fit yields tolerable results: global fit measures such as the *Comparative Fit Index* (CFI), the *Tucker-Lewis Index* (TLI), or the *Root Mean Square Error of Approximation* (RMSEA) attest an adequate fit of the model to the data. Thus, the analysis is proceeded without any further modification of the model.

Another essential prerequisite of multigroup analysis is measurement invariance.

It only makes sense to compare paths between latent variables among different groups if the latent construct is measured the same way among these groups. Different increments of measurement invariance exist, the most basic one is the invariance of factor loadings which is also called metric invariance. With regards to the model above, factor loadings between the latent variables of noise annoyance as well as physical and mental health must not vary between married and unmarried people. Otherwise, if measurement variance exists, the comparison of all other paths in the model is invalid.

Table 6.2: Model Fit for the SEM

	Value
Number of Observations, Unmarried	1383
Number of Observations, Married	1755
Estimator	MLR
χ^2	1761.817
CFI	.901
TLI	.865
RMSEA	.049

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Although the classic χ^2 difference test for a model with fixed and another model with free factor loadings reveals that measurements differ between unmarried and married people (χ^2 difference = 21.603, $p \leq 0.01$), there is some critique on using this test. For example, van de Schoot, Lugtig, and Hox argued that it performs unsatisfactory in cases of large sample sizes (van de Schoot et al., 2012, 487). Instead, other measures such as the CFI or TLI should also be used to inspect whether the models still fit the data well. As this application's sample also is large (N Unmarried = 1383; N Married = 1755) and, e.g., CFI and TLI do not significantly differ between the models, measurement invariance of factor loadings between married and unmarried people is assumed to exist.[34] Therefore, the latent variables in the analysis are assumed as being equal among those groups.

Table 6.3 contains the factor loadings of the measurement invariant latent variables of noise annoyance as well as physical and mental health. By inspecting the magnitudes of the loadings, it can be inferred that the latent variable for noise annoyance and physical health are the two best measured latent constructs. All of their manifest

34 Details of this analysis can be found in Table A.2 of the Appendix.

variables load higher than .617 and .745, respectively, on the corresponding latent construct. The mental component score shows a weaker general fit with loadings higher than .463. Generally, all three latent constructs fit the data reasonably and adequately to conduct the analysis.

Table 6.3: Loadings of Manifest Variables on the Latent Variables (N Unmarried = 1383; N Married = 1755)

	λ	SE	95% CI
Physical Component Score			
General Health	.766***	.022	[.723, .809]
Physical Functioning	.817***	.023	[.772, .862]
Role Physical	.830***	.022	[.788, .873]
Bodily Pain	.745***	.022	[.701, .789]
Mental Component Score			
Vitality	.482***	.025	[.434, .530]
Social Functioning	.831***	.033	[.767, .896]
Role Emotional	.820***	.032	[.757, .883]
Mental Health	.463***	.024	[.416, .509]
Noise Annoyance			
Annoyance During Day Time	.873***	.046	[.783, .963]
Annoyance During Night Time	.617***	.036	[.546, .688]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

6.4.2 Social Buffering of Marriage

Are married people less likely to report health problems when they are exposed to the deleterious effects of road traffic noise stressors? To answer this research question, section 6.1 conceptualized a multivariate relationship between the involved variables which contains mediation as well as moderation effects. Road traffic noise affects health directly, but also indirectly through noise annoyance. Furthermore, these effects vary between the groups of married and unmarried people because married people can draw on more social buffering resources. Table 6.4 exhibits the results of the analysis of this complex relationship.

This table contains different regressions that depict the paths in Figure 6.4. It includes the regressions for physical and mental health on noise annoyance and road

traffic noise, as well as the regressions for noise annoyance on road traffic noise. Based on these estimates, the indirect effects of road traffic noise and noise annoyance on the health variables as well as total effects on the health variables are computed.

Table 6.4: Standardized Linear Regression Coefficients of the SEM Model (N Unmarried = 1383; N Married = 1755)

	Unmarried			Married		
	β	SE	95% CI	β	SE	95% CI
Physical Component Score						
Noise Annoyance	-.114***	.026	[-.165, -.062]	-.098***	.027	[-.151, -.044]
Road Traffic Noise	-.025	.074	[-.171, .120]	.047	.090	[-.129, .223]
Mental Component Score						
Noise Annoyance	-.134***	.032	[-.197, -.071]	-.090**	.031	[-.151, -.030]
Road Traffic Noise	-.199†	.106	[-.407, .009]	.124	.099	[-.070, .318]
Noise Annoyance						
Road Traffic Noise	.196†	.103	[-.006, .399]	.413***	.089	[.239, .587]
Indirect & Total Effects						
RTN → ANNOY → PCS	-.022†	.013	[-.047, .003]	-.040**	.015	[-.069, -.011]
RTN → ANNOY → MCS	-.026†	.016	[-.058, .005]	-.037*	.015	[-.067, -.008]
Total PCS	-.048	.077	[-.198, .103]	.007	.091	[-.172, .186]
Total MCS	-.225*	.106	[-.432, -.018]	.087	.101	[-.112, .285]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); all models are controlled for age, gender, income, education, smoking, and municipality size

As expected, differences between married and unmarried people with regards to the individual paths exist, but they do not necessarily correspond to the theoretically assumed relationships. While the direct effects are similar to what may have been expected, the effect of road traffic noise on noise annoyance is stronger for married than for unmarried people. The theory of stress buffering would assume that married people perceive road traffic noise as less annoying. At the same time, however, the effects of noise annoyance on physical and mental health are still higher for unmarried people than for married ones. Generally, the direct effects of road traffic noise on health are rather small, they lack magnitude and only for unmarried people the effect of road traffic noise on mental health is statistically significant on a level of $p \leq .10$. In any case, given the difficulties of the spatial linking procedures between road traffic noise and the georeferenced survey data, the results of this analysis still yield some interesting insights.

Moreover, the indirect effects of the analysis also reveal interesting patterns and support some of the findings of the direct effects. Noise annoyance (indicated by the term ANNOY) has a differential mediating effect between unmarried and married people, both with regards to physical and mental health. Married people report worse physical and mental health if road traffic noise affects them and the more it concerns them. These effects are comparatively lower for unmarried people. Then again, unmarried people are the only group that reports worse mental health concerning the total effect of road traffic noise, which corroborates the findings from the direct path between road traffic noise and mental health.

Before the theoretical implications of these findings are discussed, it is of high importance to test whether the found differential effects are not due to statistical randomness in the first place. To this end, the following section presents invariance tests of the separate paths. These invariance tests are vital to infer any theoretical implications from the above-presented findings.

6.4.3 Robustness Check: Invariance of Paths

The paths of the SEM in this application show noticeable results and differences between the groups of married and unmarried people. Some of the differences, however, are rather small; thus, it remains unclear whether these differences are due to random variation or due to a real difference between the groups. For this reason, the following section establishes statistical invariance tests for these paths to answer the question of random variation.

Invariance tests are conducted for paths that are relevant for the theory, comprising the paths in Figure 6.4. In detail, these paths involve the direct effect of noise annoyance on physical health (ANNOY $\rightarrow$ PCS); the direct effect of noise annoyance on mental health (ANNOY $\rightarrow$ MCS); the direct effect of road traffic noise on physical health (RTN $\rightarrow$ PCS); the direct effect of road traffic noise on mental health (RTN $\rightarrow$ MCS); the direct effect of road traffic noise on noise annoyance (RTN $\rightarrow$ ANNOY); the indirect effect of road traffic noise via noise annoyance on physical health (RTN $\rightarrow$ ANNOY $\rightarrow$ PCS); the indirect effect of road traffic noise via noise annoyance on mental health (RTN $\rightarrow$ ANNOY $\rightarrow$ MCS); the total effect on physical health (TOTAL PCS); the total effect on mental health (TOTAL MCS); and a model that constrains all regression paths to be equal between married and unmarried people (ALL). To compare freely and fixed regression paths χ^2, CFI, TLI and RMSEA as well as the *Bayesian Information Criterion* (BIC) and *Akaike Information Criterion* (AIC) are used to inspect invariances, as shown in Table 6.5.

Table 6.5: Tests for Invariance of Paths Between Unmarried and Married People (N Unmarried = 1383; N Married = 1755)

	χ^2	df	p	CFI	TLI	RMSEA	BIC	AIC
ANNOY → PCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1762.498	170	0	.901	.866	.049	143213.3	142608.1
ANNOY → MCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1764.230	170	0	.901	.865	.049	143215.0	142609.9
RTN → PCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1762.270	170	0	.901	.866	.049	143213.0	142607.9
RTN → MCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1768.533	170	0	.900	.865	.049	143219.3	142614.2
RTN → ANNOY								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1765.881	170	0	.900	.865	.049	143216.7	142611.5
RTN → ANNOY → PCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1766.564	171	0	.900	.866	.049	143209.3	142610.2
RTN → ANNOY → MCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1768.279	171	0	.900	.866	.049	143211.0	142611.9
TOTAL PCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1767.088	172	0	.900	.866	.049	143201.8	142608.7
TOTAL MCS								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1775.692	172	0	.899	.866	.049	143210.4	142617.3
ALL								
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1782.652	187	0	.896	.872	.048	143096.6	142594.3

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); all models are controlled for age, gender, income, education, smoking, and municipality size

In brief, by comparing the fixed paths separately in one-to-one connections with the freely estimated ones, the model fit changes only slightly. For all models, except the one in which all regression paths are fixed, the fit measures differ non-substantially.

Because of the high number of observations, in all models the χ^2 values increase, indicating a less good fit of these constrained models. While even in the most constrained model (ALL), the RMSEA point estimate decreases by .001, the confidence intervals between the models still are overlapping (not shown in the table). Generally, fixing parameters between the groups of married and unmarried people does not yield any evidence that the differences of coefficients occur at random.

In line with the theory, it makes sense to estimate separate models for married and unmarried people. In the complex of road traffic noise, noise annoyance, as well as physical and mental health differential estimates exist. Some of the estimates still point in other directions than expected, most notably the estimates for noise annoyance. Consequently, they receive special attention in the following discussion that concludes this application.

6.5 Discussion

Married and unmarried people respond differently to external stressors, which motivated this research to ask whether this variation also exists for road traffic noise stressors and health. One mechanism is that married people can draw on more coping resources than unmarried people, and these coping resources help to reduce the deleterious effects of stressors. Another mechanism is that stressors are not perceived as stressors by married people in the first place because stress processing in the brain differs between married and unmarried people. Although both mechanisms are summarized under the umbrella of social buffers, the latter emphasizes the role of the perception of stressors. The results provide support for both mechanisms, but they are more complex than expected.

First, neither in the group of married people nor in the group of unmarried people the main effects of road traffic noise on the health indicators show decisive results. Solely in the group of unmarried people, there is an adverse effect of road traffic noise on mental health. Second, the effect of road traffic noise on noise annoyance is higher for married people than for unmarried people. One might wonder whether this leads to different indirect effects of road traffic noise on health via noise annoyance between married and unmarried people, which is corroborated by a mediation analysis: the effect is stronger for married people. Third, noise annoyance as a main effect on mental health is higher for unmarried people than for married people. This finding means that whereas married people on average respond more annoyed to road traffic noise, the actual annoyance has a higher impact on mental health for unmarried people. Overall, the sum of the direct effects of road traffic noise and its mediation

through noise annoyance only shows a significant effect for the unmarried group with regards to mental health.

6.5.1 The Changing Role of Marriage in Society

What more and more researchers emphasize is the changing role of marriage in society. Being married is no longer a necessity for living with a partner, and long-term relationships without a marriage certificate are increasingly common (Liu & Reczek, 2012, 794). Instead, it was argued that "legal marital status no longer reflects today's social reality" (Drefahl, 2012, 463). Indeed, differences between societies exist, but particularly the German case presented here shows a development towards more unmarried arrangements after the German reunion in 1990 (Statistische Ämter des Bundes und der Länder, 2015, 30). Accordingly, with this changing role of marriage also comes a new definition of its social buffers.

As roles are changing, researchers are confronted with a moving target. Studies on social buffers and marriage that have been conducted decades ago may no longer apply to contemporary studies. Differences between married and unmarried people vanished over the years, and this may explain why this application only found small differences between these two groups. While married people generally still live healthier lives than unmarried people (Carr & Springer, 2010), the mechanisms that lead to such differences may have changed (see, e.g., Lin, Chen, & Li, 2016 or Johnson, Horne, & Galovan, 2016). Studies on social buffers and marriage should reflect on the specific societal setting of their respondents.

Future research, moreover, could differentiate between the group of married people and all people who belong to the group of unmarried people. This group involves singles, people in romantic relationships as well as those who are separated, divorced or widowed. In this application, such differentiation was not possible to conduct due to small sample sizes in these groups. Specialized studies dealing with the topic of the family are more suitable for this purpose (Brüderl et al., 2015; Schmiedeberg, 2015). However, in this application, the group of married people still was the healthiest group, but researchers should be cautious with regards to confounding group characteristics of the other groups.

6.5.2 Noise Measurements Data in Survey Research

However, such small effects are interesting in comparison to other studies that use data on specific groups in specific geographic regions (Bocquier et al., 2013). These

studies, in contrast, showed stronger effects, e.g., of air traffic noise on health (Boes et al., 2013). Moreover, they often used more observational operationalizations of health such as blood pressure (Babisch, 2014) or blood sugar measures (Eriksson et al., 2014). It remains unclear whether subjective measurement methods are not suitable for capturing the effects of noise stressors on health. At the same time, other studies corroborated the validity of subjective measures in large-scale survey research (Hank, Jürges, & Schaan, 2009). Further research with observational health measures in a large-scale survey would help to find hints to solve this puzzle.

6.5.3 Conclusion

The effects found in this study are small, and it is questionable whether they qualify at all as substantial effects. Other factors, such as age or education, are estimated more precisely (see Table A.3 of the Appendix). Comparable to other studies on health differences in society, these effects are in-person differences—personal characteristics explain more variation in health than any other contextual factor. In comparison to this extensive body of research, the findings of this application indeed are minor.

Nevertheless, mainly because so many factors influence people's health, it is interesting that road traffic noise stressors show at least some effects in this application. These effects corroborate theoretical considerations with regards to the health effects of marriage by introducing external evidence via traffic noise measurements. They were not self-reported by the survey respondents but were gathered from auxiliary geospatial data sources. From this perspective, the enrichment of survey data with external sources leads to the validation of existing survey measurements.

This empirical application, thus, exemplifies the use of georeferenced survey data for existing and new research questions. It combines two lines of research—family research and environmental stress research—through the method of spatial linking of road traffic noise data to survey respondents' locations. This method makes it possible to question existing findings in family research while also providing new evidence for stress research. Some theoretical and methodological concerns remain as both data sources of this application were not collected primarily to combine them. An instrumental variable approach has to be applied to create a representative sample. Future survey studies should invest in finding new geospatial data sources that researchers can use equally for all survey respondents in a study.

7 Application II: Ethnic Diversity and Xenophobia

This empirical application is built on a cooperation and joint work of me and my colleagues Julia Klinger and Merlin Schaeffer. It was published as a German research article in the German sociological journal "Zeitschrift für Soziologie" (Klinger et al., 2017).[35] While the research question is similar in this application, the elaborations on the methodological background and its incorporation into the issues of using georeferenced survey data go beyond the original scope of the article. Thus, this empirical application presents a contribution to the field that would have exceeded the scope of the article.

7.1 Research Question

The social sciences have provided a large number of theories and hypotheses to explain prejudices and xenophobia (Allport, 1954; Blalock, 1967; Blumer, 1958). Because these theories generally defined xenophobia as fear or threat of people who perceive others as foreign, usually they emphasized the role of social interaction between groups of people. Specifically, the question arose whether the amount of foreign people in neighborhoods increases or decreases xenophobia (Quillian, 1995; Schneider, 2008). In recent years, Western societies such as Germany have witnessed an increase in the discussions about the relationship between migration and xenophobia. Accordingly, theories that explain these developments of xenophobia are still of high interest, not just for the social sciences but also for societies as a whole.

Despite a rather long research tradition, the actual mechanisms of social interaction are not yet clarified. For sure, norms, values, and other traits have a high influence on people's perception of foreigners, such as immigrants (Raijman, Davidov, Schmidt, & Hochman, 2008). At the same time, researchers repeatedly have empha-

35 My main role in this article was to transfer requirements of theoretical considerations to the empirical operationalizations of the Halo constellations. I introduced the GIS tool of focal linking as a procedure to relate direct and surrounding neighborhoods to each other and provided solutions to analyze and summarize the results. Moreover, I implemented all steps of data processing and the final analysis as a reproducible software routine which can be accessed here: https://datorium.gesis.org/xmlui/handle/10.7802/1493?locale-attribute=en. Generally, we agreed that each author of the article contributed to the article as much as any other, hence the alphabetical order of the authors. All the authors gave me their consent to re-use our findings. Also, all figures and tables in this empirical application are not in the form as they were published in the article. The publisher, Walter de Gruyter GmbH, confirmed to me that nothing stands in the way of such a form of reproduction.

sized the influence of contextual determinants that shape people's attitudes (Janmaat, 2014; Schlueter & Scheepers, 2010; Semyonov, Raijman, Tov, & Schmidt, 2004). Often, they understand contextual determinants as a spatial dimension, more specifically as neighborhoods that affect people in their everyday lives (Sharkey & Faber, 2014). In this regard, neighborhoods are a place of social interaction and xenophobia directly relates to their composition and their geographic size.

This application presents a new approach in the social sciences that emphasizes the role of neighborhood constellations (Gleditsch & Weidmann, 2012). It originates in the idea that while neighborhoods as single entities influence people's lives, they structurally and spatially also relate to other neighborhoods which, in turn, influence people as well (Legewie & Schaeffer, 2016). Not only direct neighborhoods of people affect xenophobic attitudes through social interaction but also surrounding neighborhoods can play a role in this interplay. The following section starts with presenting the most common theories for xenophobia and social interaction. It then turns to the new approach that emerged under the term of the *Halo Hypothesis*. Finally, this theoretical part closes with a discussion of the different geographic sizes of neighborhoods that were used in previous research and which are relevant here.

7.1.1 Contact, Intergroup Threat and the Ethnic Competition Theory

Social science scholars commonly use two distinct sets of theories to explain prejudices and, in consequence, xenophobia related to neighborhoods: the classic *Contact Theory* (Allport, 1954) or the *Intergroup Threat Theory* (Stephan et al., 2009) which is similar to the *Ethnic Competition Theory* (Banton, 1983; Blalock, 1967; Blumer, 1958). Both of these sets of theories discuss xenophobia with regards to social interaction between people in neighborhoods. They emphasize the ethnic diversity of neighborhoods and how the presence of people who are perceived as foreign shape attitudes towards them. While both sets use similar concepts for their theories, the potential outcomes relating to xenophobia are in stark contrast.

Allport's Contact Theory stresses the significance of contact between in-group and out-group people and its potential role in minimizing prejudices. To reduce xenophobia, contact between people though must not be "casual", or as Putnam puts it "superficial" (Putnam, 2007), as "such contact does not dispel prejudice; more probable it increases it" (Allport, 1954, 251). When people see or observe other people without knowing their motives and why they act in certain ways, they probably develop negative attitudes. Only if people establish non-casual contacts and develop positive interactions, they can reduce prejudices and develop social-emphatic proximity to each

other (Pettigrew, 1998; Pettigrew & Tropp, 2006). People gain more information about the out-group, learn about their norms and values, and establish affective bonds that help to limit misunderstandings. Generally, spatial proximity increases the likelihood of positive interaction (Weins, 2011), and thus, it is assumed that higher rates of immigrants in neighborhoods also decrease xenophobia.

The second set of theories, the social-psychological Intergroup Threat Theory or the sociological Ethnic Competition Theory, comes to different conclusions with regards to contact between in-group and out-group people and its influence on prejudices. Albeit comparable to the Contact Theory and its restriction on casual contact, these theories stress the role of potential feelings of threat, e.g., when a relative change of foreign people occurs in a neighborhood (Hopkins, 2010). These feelings can be rooted in economic fears, such as job loss, price collapses of housing prices or diminishing of general wealth; or they can be symbolic when people are afraid to lose their cultural dominance. Generally, because these fears affect economic and symbolic domains, researchers assume that specific sociodemographic groups, for example, low-educated and unemployed people (Helbling, 2011; Hello, Scheepers, & Sleegers, 2006), homeowners (Nowak & Sayago-Gomez, 2018), and conservative people (Raijman et al., 2008) are more vulnerable to such fears. Typically, both of these theories assume that the spatial proximity of people increases threats and competition, and thus, it is assumed that higher rates of immigrants in neighborhoods increase xenophobia.

Both sets of theories have been successfully applied in research (Quillian, 1995; Schneider, 2008; Semyonov, Raijman, & Gorodzeisky, 2006; Weins, 2011). While indeed both theories provide plausible and convincing arguments, as outlined above, it is confusing that their competing hypotheses with regards to immigrant rates should apply simultaneously. A crucial role plays the concept of spatial proximity and the question of whether it increases the likelihood of positive interaction between people (Schlueter & Scheepers, 2010). When people have the opportunity to convert spatial proximity into social proximity, positive interaction can occur. Newer and more fine-grained theories, therefore, concentrate on the specific spatial conditions of neighborhoods to disentangle spatial proximity from social proximity (Martig & Bernauer, 2016; Teney, 2012; Weßling, Hartung, & Hillmert, 2015). One of the hypotheses that provide a reconciliation between the competing theories is the Halo Hypothesis which the following section presents.

7.1.2 The Halo Effect Hypothesis

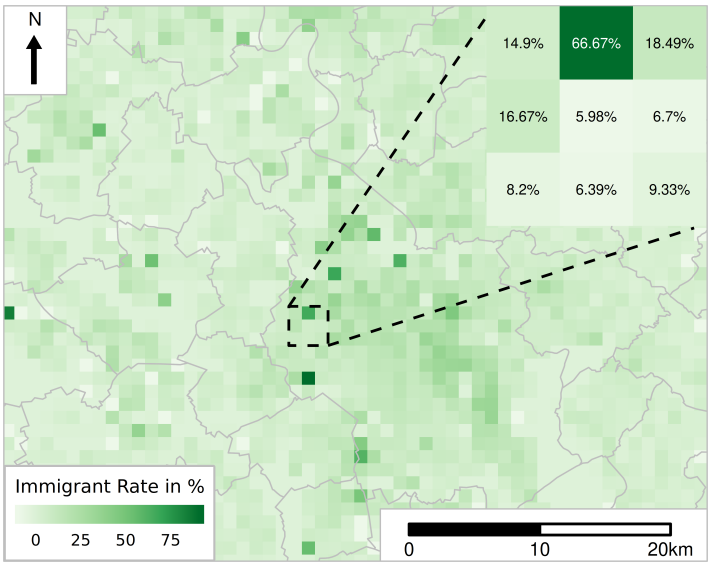
The Halo Hypothesis' most striking lesson is that the above-presented theories are only apparently incompatible. Instead, supporting findings for each one of them in different applications occur because the measurement of neighborhoods' immigrant rates may have been collected on different levels or types of neighborhoods. Some applications used measures on a small neighborhood level, others on the city level. Further studies even used measures on the administrative district levels so that comparisons between each one of these applications would only be valid if immigrant rates did not change between different geographic levels. According to the MAUP (see Chapter 3.2.2), this is not a realistic assumption. Differences, as well as similar predictions of the theories, may be caused by unsynthesized neighborhood operationalizations.

What the Halo Hypothesis proposes is to differentiate between direct neighborhoods and surrounding neighborhoods (Bowyer, 2008; Rydgren & Ruth, 2013). Direct neighborhoods are the types of neighborhoods with day-to-day interaction, where people meet other people of their neighborhood, and positive social interactions are likely to occur (Sharkey & Faber, 2014). Surrounding neighborhoods are seldom frequented by people, and positive social interaction is less probable. They are always less familiar, and in cases of varying ethnic composition between direct and surrounding neighborhoods, the latter may be perceived as even more different. Thus, the difference between direct and surrounding neighborhoods is dynamic and depends on the ethnic composition of each one of them.

This dependence on the constellations between direct and surrounding neighborhoods is the reason why both Contact Theory and Intergroup Threat/Ethnic Competition Theory come to compliant predictions under certain circumstances. For example, when the ethnic composition between neighborhoods does not vary, borders between them may not be perceived as manifest borders. In contrast, when the ethnic composition varies, borders may be noticed as reasonably manifest. Thus, when surrounding neighborhoods have higher rates of immigrants, Contact Theory and Intergroup Threat/Ethnic Competition Theory both would predict higher likelihoods of prejudices and hence xenophobia. The Intergroup Threat/Ethnic Competition Theory would predict this link between immigrant rates and xenophobia in any case, while the Contact Theory would do it because of the likelihood of sole casual contacts in surrounding neighborhoods. The interplay of direct and surrounding neighborhoods with regards to ethnic composition forms social boundaries that combine predictions from Contact Theory and Intergroup Threat/Ethnic Competition Theory. That is the core of the Halo Hypothesis.

What gives the Halo Hypothesis its name is that such neighborhood constellations

are merely relationships of neighborhoods in a circular form. It is the potential interplay of different levels of ethnic diversity that can lead to the xenophobia of people (Rydgren & Ruth, 2013). As such, the Halo Hypothesis combines elements of the two hypotheses and furthermore of the *Social Isolation Theory* (Putnam, 2007, 141ff): direct neighborhoods build islands in a network of neighborhoods when they differ severely from the surrounding neighborhoods (see also for another context Lim, Metzler, & Bar-Yam, 2007). Figure 7.1 exemplifies such a constellation for a 1 km^2 census neighborhood in the city of Cologne. While the direct neighborhood (the inner 1 km^2 grid cell) comprises a comparably low immigrant rate, some of its surrounding neighborhoods (the surrounding 1 km^2 grid cells) comprise far larger immigrant rates. Direct neighborhoods that are surrounded by neighborhoods with diverging rates of immigrants constitute archetypal examples of Halo constellations.



Data Sources: Statistical Offices of the Federation and the Länder (2016) and Federal Agency for Cartography and Geodesy (2018)

Figure 7.1: Halo Constellation of Immigrant Rates Between Direct and Surrounding Neighborhoods in the City of Cologne

In their day-to-day life, people from such direct neighborhoods may have tentatively low opportunities to meet foreign people. Surrounding neighborhoods are of peripheral importance so that people also have low opportunities to interact with people from these neighborhoods. What is more, if these people from the surrounding neighborhoods are foreign and have different ways of arranging their everyday life, people from direct neighborhoods may feel threatened by their presence. People from direct

neighborhoods merely do not know much about surrounding neighborhood people's life, and as such develop prejudices and reservations. In this specific form of a Halo constellation, people are at risk to develop considerable levels of xenophobia.

Authors with similar hypotheses in other research fields already found evidence that supports conceptualizing neighborhoods in this direct-surrounding-distinction. Some of the most prominent examples are prohibition policies against Catholic immigrants in the 19th century in the United States of America (Andrews & Seguin, 2015) and election results of right-populist parties in England (Bowyer, 2008), Sweden (Rydgren & Ruth, 2013) or Switzerland (Martig & Bernauer, 2016). With regards to the German societal context though it is not yet clear how such constellations affect attitudes towards immigrants. In comparison to other Western societies, ethnic segregation and ethnic diversity between neighborhoods is comparably low (Schönwälder & Söhn, 2009). Even so, in sum, this empirical application starts with the hypothesis that also in Germany xenophobia systematically increases in direct neighborhoods which have lower levels of ethnic diversity in comparison to their surrounding neighborhoods.

7.2 Data and Measures

7.2.1 Sample: GGSS 2014 and German Census 2011

The data of this empirical application stem from the combined data of the georeferenced GGSS 2014 (see Chapter 2.3.1) and geospatial data from the German Census 2011 (see Chapter 2.4.2). Spatial linking methods enable to add auxiliary attributes of the German Census 2011 data to the GGSS 2014 data to create measures for the Halo constellations. In sum, the sample data consist of 1,192 survey respondents who also have the German citizenship.

The following section first describes in detail how both data sources are combined using spatial linking methods. As such, all geospatial data measures are introduced, before subsequently all survey data measures are presented. A table of descriptives that gives an overview of the whole set of variables can be found in Table 7.2.

7.2.2 Geospatial Data Measures

Geospatial Data Preparations

The data of the German Census 2011 for the whole extent of Germany are publicly available as a Comma Separated Values (CSV) dataset and can be downloaded from

the internet. Despite their similarities to other data from EU regulations or directives (see Chapter 4.1.1), users can directly access the data. As introduced in Chapter 2.4.2, these data contain information for the sociodemographics of the German population in uniformly shaped grid cells.

Moreover, because the data are already harmonized, preparing them for analyses is straightforward. Each row of the data is a single centroid geo-coordinate out of the 557,256 raster cells in Germany and refers to a raster cell resolution of 1 km $\times$ 1 km. By converting them to a raster data file and assigning an CRS (see Chapter 2.4.1), the data are convenient to use.[36] The resulting file can be processed and analyzed in an GIS and provides information on the population of the whole extent of Germany.

Focal Linking

The Halo Hypothesis assumes relationships in specific neighborhood constellations of the direct and surrounding neighborhoods. Accordingly, it makes no sense to use spatial linking by location methods to add census attributes to the survey respondents' geo-coordinates because the information about the neighborhood relationships would be lost. Instead, the neighborhood constellations have to be derived from the data in a way that relates the direct neighborhoods to the surrounding neighborhoods. Only after operationalizing the neighborhood constellations in such a way, this information can be spatially linked to the respondents' locations.

What is a useful method to create the neighborhood constellations is focal linking, described in Chapter 5.2.3. In detail, applying the method to the demands of the research question results in a three-step procedure:

1. For each raster cell in the data, a focal operation is conducted using neighborhood matrices.

2. The new raster cell values, containing the results of the focal operation, are projected into a joined coordinate space with the survey respondents' locations.

3. Both data sources are spatially linked in such a way that the survey respondents' data contain new information about the results of the focal operation, which is their neighborhood constellation.

---

36 The *R* package *georefum* (https://github.com/stefmue/georefum) already contains prepared raster data for the German Census 2011 that were created using the described approach. It also provides the routines for creating the files—details on their application can be found in Müller* et al. (2017).

The idea of the Halo Hypothesis leaves room for different combinations to apply the focal linking method. Generally, what researchers who test the theory aim to do is to relate the ethnic diversity of the direct neighborhood to the ethnic diversity of the surrounding neighborhoods. Using 1 km^2 grid cells for the direct neighborhood, for example, is just one possibility though, other researchers also used spatial units on a larger geographic scale (Martig & Bernauer, 2016). Creating measures of neighborhood constellations should take into account discussions on (a) the geographic scale of direct and surrounding neighborhoods, (b) descriptive statistics used for creating measures of neighborhoods, and (c) how direct and surrounding neighborhoods can be related for the actual analysis.

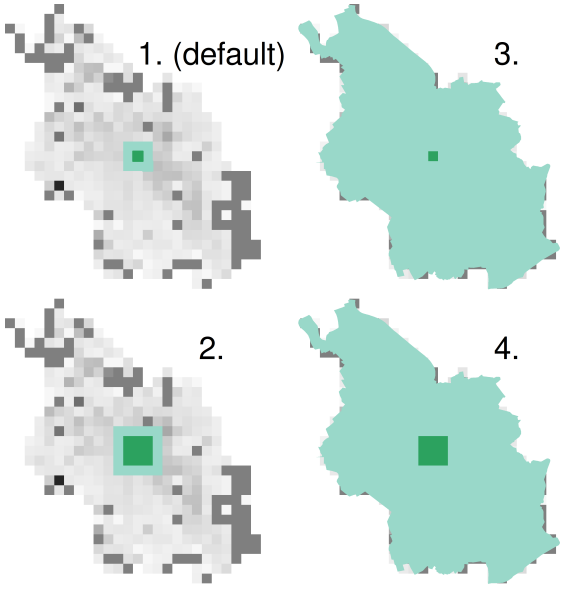
As mentioned, the (a) geographic scale of neighborhoods affects direct and surrounding neighborhoods at the same time. While this application uses the 1 km^2 grid cells as default operationalization for direct neighborhoods, there may also be reasons for a broader range of grid cells. For example, the "effective neighborhood" (Spielman & Yoo, 2009, 1100), as the interaction radius of people in their day-to-day routine, is often larger. People commute to work or they go shopping far away from their home. Therefore, this application also uses broader direct neighborhoods of the size of 9 km^2 as an alternative. The same argument applies to the surrounding neighborhoods because they could be even bigger in comparison to effective neighborhoods. Besides its default operationalization of the adjacent eight 1 km^2 grid cells as surrounding neighborhoods, this application uses the adjacent sixteen 1 km^2 grid cells as surrounding neighborhoods for the alternative operationalization of the direct neighborhoods of 9 km^2. Moreover, as another alternative for the surrounding neighborhoods, the whole area of the municipalities is introduced, both for the direct neighborhood with 1 km^2 and the 9 km^2 grid cells.

Table 7.1: Geographic Size Combinations of the Plausible Halo Constellations

	Direct Neighborhood	Surrounding Neighborhood	Used as...
1.	1 km^2 Grid Cell	Adjacent Eight 1 km^2 Grid	default
2.	9 km^2 Grid Cell	Adjacent Sixteen 1 km^2 Grid	alternative
3.	1 km^2 Grid Cell	Surrounding Municipality	alternative
4.	9 km^2 Grid Cell	Surrounding Municipality	alternative

Table 7.1 presents each of these options in the corresponding combinations. Moreover, for a more visual presentation, Figure 7.2 displays these combinations in a map of the city of Cologne. While the default operationalization—1 km^2 for the direct neighborhood and the adjacent eight 1 km^2 grid cells for the surrounding neighborhood—represents the Halo constellation on the smallest geographic scale,

the areas covered by all other Halo constellations increase with the choice of neighborhood sizes. Generally, the entire set of these Halo constellations provides plausible operationalizations, at least as the Halo Hypothesis is not yet clear in that regard. Thus, it is also an empirical question of whether the results of the analyses with different Halo constellations differ.



Data Sources: Statistical Offices of the Federation and the Länder (2016) and Federal Agency for Cartography and Geodesy (2018)

Figure 7.2: Four Different Types of Plausible Halo Constellations

Apart from the geographic size, the question remains which (b) descriptive statistic may be appropriate to reflect people's perception of their surrounding neighborhoods. Do people experience attributes of their neighborhood according to the average, or do they orient towards extreme values? Again, this question cannot be answered from a theoretical standpoint. Therefore, this application uses both the mean value and the maximum value for all operationalizations of the surrounding neighborhoods. The mean value represents a more aggregate measure that decreases variance. The maximum value, on the other hand, depicts a more nuanced measure that is sensitive to variance, which qualifies it as the default descriptive statistic. Effectively, the mean value and the maximum value serve as two competing measures of the surrounding neighborhoods.

Finally, (c) different ways of relating the direct neighborhood and the surrounding neighborhood measures to each other exist. The first one is to compute differences between the surrounding neighborhoods and the direct neighborhoods. In such a

measure, the value of 0 depicts a turning point of the scale—values above this limit are, by definition, archetypal Halo constellations. All differences, however, are the same across the whole range of the immigrant rate scale, no matter if the difference originates from the values of 20% and 5% (=15) or 35% and 20% (=15). A second and alternative measure is to compute ratios. In this measure, the value of 1 depicts a turning point. While this measure is bound to a value of 0, its advantage is that a ratio results in different values for the input values of 20% and 5% (=4) and 35% and 20% (=1.75). For this reason, this measure is the default measure to relate direct and surrounding neighborhoods in this application. One last option is to leave the direct and surrounding neighborhoods as they are and to include them as an interaction term in regression models. Then again, if this model was a linear model, the interaction would not respect different slopes at different scale locations.

7.2.3 Survey Data Measures

Xenophobia

The GGSS 2014 comprises an additional questionnaire with items that ask for attitudes towards immigrants and people with an immigration history. Respondents rated on a scale of 1 (”Agree strongly”) to 5 (”Disagree strongly”) whether they agree to the following statements: ”Immigrants increase crime rates”; ”Immigrants are generally good for Germany’s economy”; ”Immigrants take jobs away from people who were born in Germany”; ”Immigrants improve German society by bringing new ideas and cultures”; and ”Germany’s culture is generally undermined by immigrants”.

By combining these items on one scale, they qualify to measure xenophobic attitudes of the respondents as an PCA corroborates. While the positive and negative worded items show a distinct method’s effect—i.e., the analysis reveals a principal component for the positive items and one for the negative ones—, the extraction of one single principal component of the z-standardized items with varimax rotation still explains 57% of the variance. Accordingly, it can be assumed that such a single component solution yields an appropriate approximation of the items’ input matrix (details of this analysis can be found in Table B.1 of the Appendix). The values of the dependent variable xenophobia are the corresponding factor scores of the component.

Sociodemographics

Not just the contextual factors presented above are hypothesized to determine xenophobia. Instead, other sociodemographic factors influence how people express their attitudes towards immigrants. To adjust the results for these factors, the analysis contains additional control variables for the respondents' age (in years), gender, education ("low": ISCED 1-2, "medium": ISCED 3, "advanced": ISCED 4-5 and "high": ISCED 6-8), net income (in EUR), and binary indicators for unemployment and homeownership. The latter aims to measure differences in the personal significance of the neighborhood and serves as a proxy for the duration of residence in the neighborhood.

Contextual Controls

Finally, it is controlled for variables that may have an effect on the estimates. These variables comprise a binary indicator for the region of Germany in which respondents live (Western Germany: 0, Eastern Germany: 1), the size of their municipality ("Rural Community": below 5,000 inhabitants, "Small Town": 5,000 to 19,999 inhabitants, "City": 20,000 to 99,999 inhabitants, "Big City": at least 100,000 inhabitants), the number of inhabitants in their direct neighborhood as well as the mean flat size in square meters per inhabitant in their direct neighborhood. While most of these variables adjust regional factors of people's living arrangements, the mean flat size can serve as a proxy measure for wealth in the neighborhood. Table 7.2 gives an overview of all variables used in the default analysis, including the default Halo operationalization and the individual level variables.

Table 7.2: Descriptive Statistics and Overview of All Variables of the Default Analysis (Pairwise Deletion)

	Mean / %	SD	Minimum	Maximum
Dependent Variables				
Threat	0	1.01	-2.27	2.85
Independent Variables				
Halo Ratio*	2.09	2.33		
Immigrant Share[a]*	5.71	6.32		
Maximum Immigrant Share[b]*	5.15	5.91		
Control Variables				
Gender (female)	48.91	50.01		
Age	48.96	17.97	18	91
Education				
Low	9.04			
Medium	48.48			
Advanced	18.58			
High	23.90			
Income	1460.52	1102.23	0	8750
Unemployment	5.12	22.04		
Eastern Germany	32.72	46.94		
Municipality Size				
Rural Community	25.92			
Small Town	26.17			
City	21.81			
Big City	26.09			
Home Ownership	58.44	49.30		
Inhabitants[a]*	2395.50	3073.90		
Mean Flat Size per Inhabitant[a]*	42.49	4.96		
Number of Observations	1192			

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); [a] direct neighborhood; [b] surrounding neighborhood; * some values removed due to data protection

7.3 Analysis Strategy: Robustness Through Several Estimation Methods

7.3.1 Spatial Dependencies

This application hypothesizes that political attitudes (i.e., xenophobia) are dependent on spatial attributes. If survey respondents live in proximity to each other, these political attitudes are at risk of being spatially clustered or spatially dependent (see Chapter 5.3). Respondents who live close to each other might share more similar political attitudes than those who live far away from each other. To prevent the results from being biased, spatial dependence should be tested and adjusted in the analyses.

A common approach is to search for shared variance of the dependent variables between respondents who live close to each other. If the test turns out to be positive (i.e., spatial autocorrelation exists), for example, Spatial Lag Y Models help to adjust for spatial clustering (see Chapter 5.3.2). The values of respondents being neighbors are weighted and included as autoregressive terms in the regression equation. Thus, the analysis does no longer violate the independence assumption of ordinary regression models and provides unbiased estimates.

Testing spatial correlation requires modeling spatial dependencies between observations in the data and connectivity matrices (see Chapter 5.3.2) offer the tools to do so. In the context of this application, it was decided to compute geodesic distance measures between respondents in 10 meters steps divided by their inverse. Every respondent is neighbor to any other respondent, but respondents who live far apart do have a negligibly low influence on each other. At the same time, by using 10 meters steps, this influence does not decay as fast as in plain inverse functions, a possibly more realistic assumption of neighborhood influences.

Moreover, the connectivity matrix calculations are based on the coarsened geo-coordinates of the respondents. After georeferencing the data of the GGSS 2014, data protection legislation required to extinguish the original addresses and geo-coordinates of the survey respondents. The research data center of the GGSS, however, managed to preserve geo-coordinates that are the centroids of the 1 km^2 grid cells in which respondents live. These centroid geo-coordinates are no longer personal information, but they still provide small-scale spatial information. As this application does not place importance on connections between respondents based on small distances, they also provide a reasonable approximation of an analysis with the original geo-coordinates.[37]

37 The analyses in the journal article of this application were conducted at a time where access

Also with the coarsened geo-coordinates the analysis of spatial autocorrelation yields a statistically significant amount of the dependent variable of xenophobia. Moran's I and Lagrange Multiplier Diagnostics (Bivand et al., 2008, 258ff) show small but significant indications for spatial autocorrelation. Accordingly, the analysis should control these spatial dependencies between the respondents.

7.3.2 More Sources of Unobserved Heterogeneity

An obvious influence on people's political attitudes is the result of policies. They affect how people perceive and learn about changes in the public sphere, and they signal how politics respond to such changes. Moreover, policies often occur on administrative levels, such as municipalities. For example, regional governments decide where refugee shelters are located.[38] Some authors argued that this decision impacts attitudes toward immigrants (Förster, 2018). The actual development of attitudes is, thus, affected by decisions that have been made on geographically larger levels. In consequence, people perceive their direct living environment not only as being influenced by their direct neighbors but also by developments happening in their municipality.

Researchers cannot control or even observe all these developments in their analysis. However, such factors influence model estimations and produce biases because they contain sources of unobserved heterogeneity. If the observational unit of these sources is known, e.g., the municipality level, at least researchers can adjust model estimates for such biases. A common approach is to use regression models with fixed effects for the group level (see Chapter 5.3.1). Accordingly, to control the bias of unobserved heterogeneity in the estimates of political attitudes, the use of FE models is promising.

7.3.3 Three Choices of Estimators

Because of the notes of the preceding section, this application uses three different estimators in its analysis:

1. *Spatial Lag Y Models* to adjust spatial dependencies among the survey respondents as the default model
2. *Municipality Level Fixed Effects Models* to adjust unobserved heterogeneity between

---

to the original geo-coordinates was still warranted. The results changed only slightly and insignificantly.

38 See, e.g., https://recht.nrw.de/lmi/owa/br_text_anzeigen?v_id=10000000000000000407

municipalities

3. *OLS Regression Models* to compare the models in a more convenient way

The Spatial Lag Y Model (1.) is further specified as an OLS regression model that contains an autocovariate. Generally, this model does not differ from the Spatial Lag Y Model presented in Chapter 5.3.2. While the latter estimates its autoregressive parameters within the general model estimation, in an autocovariate model, the corresponding lagged variable is calculated beforehand. This step was necessary because this application uses multiple imputed data (see below), and for Spatial Econometric Models, Rubin's Rule is not yet defined.[39] The formula of the autocovariate is as follows:

$$autocov_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}, \tag{7.1}$$

where y_j depicts the value of the dependent variable of observation j and w_{ij} the neighborhood connection between observation i and j (see Chapter 5.3.2), based on inverse weighted geodesic distances between the respondents in 10 meters steps. The value of the autocovariate $autocov_i$, therefore, is a weighted sum of the neighbors' lagged dependent variable of observation i. Its corresponding regression coefficient, represented as ρ in the regression table, signals the amount of spatial autocorrelation between respondents and controls for the potential biases of spatial autocorrelation.

Using Municipality Level Fixed Effects Models (3.) does not prevent the results from being biased because of their dependence on a small spatial scale, but they allow excluding unobserved heterogeneity between municipalities. Moreover, they do not contain any autocovariate because this, in turn, can result in a systematic error of estimates (Angrist & Pischke, 2009, Chapter 5.3), and controlling for municipality level invariant variables is not possible anymore. However, besides re-validating the results with OLS models (3.), FE models complement the estimates of the default Spatial Lag Y Model.

In sum, combining all Halo operationalizations and all estimators results in 48 different models. Because it is inconvenient to discuss their results in detail, this application uses one specific combination as default model: the Spatial Lag Y Model with 1 km^2 grid cell values as direct neighborhood and the maximum value of the eight adjacent 1 km^2 grid cells as the surrounding neighborhood. Furthermore, as a link

39 Spatial Lag Y Models with simultaneous estimation algorithms on the list-wise deleted data yielded similar results as the approach used here. The results can be found in Table B.2 of the Appendix.

between the different neighborhoods, the ratio is chosen.

Lastly, the multivariate analysis is based on 16 datasets computed with multiple imputation. These datasets are created using *Sequential Regression Imputation* (Raghunathan, 2016, 67f) with all variables on the respondents' level as input data. Alternative estimations with list-wise deletions corroborate this approach and yield similar results.

7.4 Results

Are people who live in homogeneous neighborhoods that border on ethnically diverse surrounding neighborhoods (or are even encircled by them) more xenophobic? The analysis begins with the results of the model that includes the default Halo operationalization (1 km^2 for the direct neighborhood and the adjacent eight 1 km^2 grid cells for the surrounding neighborhood), the default relation (ratio), and the default estimation procedure (Spatial Lag Y), which is shown in Table 7.3. As a first step, model 1 includes the bivariate relationship between the Halo constellations and the dependent variable xenophobia as the baseline model. Model 2 proceeds with adding all other control variables, and model 3 tests for regression discontinuity around the threshold value of 1. The latter estimation ensures testing whether the theoretical relationship also applies to direct neighborhoods that are more diverse than their surrounding ones.[40]

The results of all three models can be summarized briefly: in none of these models, an effect of the Halo constellation predictor on xenophobia is found. Neither in the bivariate nor the multivariate or the discontinuity model a substantive effect is prevalent. The Halo constellation predictor does not show any sign of substantial significance, although at least in the bivariate model it is statistically significant on a level of $p \leq .15$. In sum, no evidence for a Halo effect with regards to the default operationalization of the Halo constellations exists.

Even so, the results of the models are reliable. Given the estimates of all other predictors, the quality of the measurement of the dependent variable can be corroborated—for example, education and unemployment point to the theoretically expected direction. Moreover, the estimates for the autocovariate ρ justify including

40 Regression discontinuity is tested by creating a dummy variable that is 0 for all Halo constellations of ≥ 1 and 1 for all Halo constellations of < 1. The interaction of this variable with the Halo constellation predictor tests the discontinuity around this threshold. To enable this test its value is not standardized at its mean value but a value of 1.

Table 7.3: Standardized Coefficients of Spatial Lag Y Regression Model for the Default Halo Operationalization (N = 1,192)

	Bivariate			Multivariate			Discontinuity		
	β	SE	CI	β	SE	CI	β	SE	CI
Halo	-0.046	0.029	[-0.103, 0.012]	-0.034	0.029	[-0.090, 0.022]	-0.029	0.030	[-0.087, 0.030]
Halo ≥ 1							0.002	0.100	[-0.194, 0.197]
Halo $\times$ Halo ≥ 1							-0.496	0.763	[-1.991, 0.998]
Threat ρ	0.264***	0.032	[0.201, 0.328]	0.168***	0.036	[0.098, 0.238]	0.167***	0.036	[0.097, 0.237]
Inhabitants				-0.078*	0.038	[-0.152, -0.004]	-0.076*	0.038	[-0.150, -0.002]
Flat Size				-0.007	0.036	[-0.078, 0.064]	-0.007	0.036	[-0.077, 0.064]
Municipality Size									
Rural Community (*Ref.*)									
Small Town				0.075	0.094	[-0.109, 0.259]	0.080	0.094	[-0.104, 0.264]
City				-0.020	0.098	[-0.212, 0.173]	-0.010	0.099	[-0.204, 0.185]
Big City				-0.057	0.114	[-0.280, 0.167]	-0.045	0.115	[-0.270, 0.179]
Age				0.049	0.033	[-0.016, 0.113]	0.048	0.033	[-0.017, 0.113]
Gender (female)				-0.101	0.067	[-0.232, 0.031]	-0.095	0.067	[-0.226, 0.036]
Education									
Low (*Ref.*)									
Medium				-0.154	0.110	[-0.368, 0.061]	-0.161	0.110	[-0.377, 0.055]
Advanced				-0.533***	0.124	[-0.775, -0.290]	-0.558***	0.124	[-0.801, -0.315]
High				-0.826***	0.126	[-1.073, -0.579]	-0.834***	0.126	[-1.082, -0.587]
Income				-0.023	0.038	[-0.097, 0.052]	-0.025	0.038	[-0.099, 0.050]
Unemployment				0.464***	0.140	[0.190, 0.737]	0.456**	0.140	[0.182, 0.730]
Eastern Germany				0.128†	0.074	[-0.017, 0.273]	0.128†	0.074	[-0.018, 0.274]
Homeownership				-0.086	0.069	[-0.222, 0.049]	-0.088	0.069	[-0.223, 0.047]
Intercept	0.013	0.031	[-0.048, 0.075]	0.421**	0.136	[0.155, 0.687]	0.427**	0.138	[0.156, 0.699]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

it in the model as it remains high and statistically significant in all models. Accordingly, it is not misspecification of the models that causes non-existing effects of the Halo constellations.

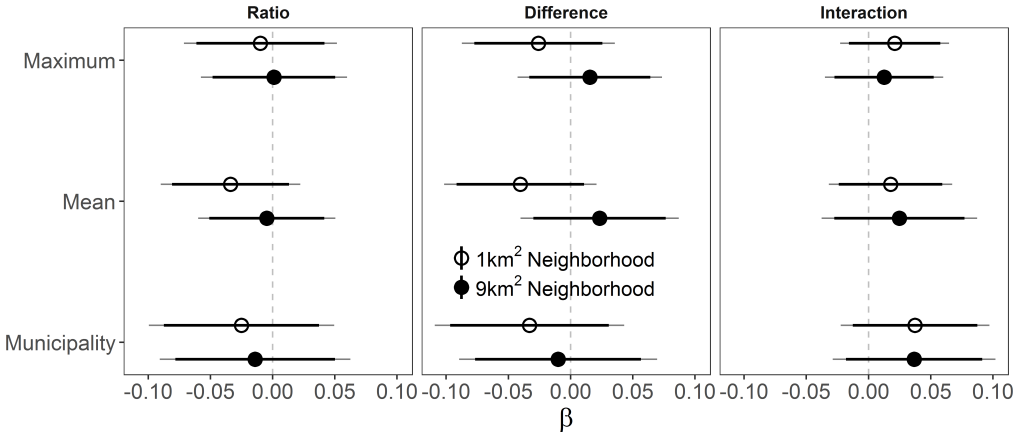
As already discussed what makes the analysis demanding are the different possibilities of operationalizing the Halo constellations. Different descriptive statistics for the surrounding neighborhoods exist (mean or maximum values), as well as different relations between direct and surrounding neighborhoods (ratios, differences, and interactions), and the choice of different geographic sizes (see Table 7.1). Moreover, it should be ensured that different estimation procedures do not lead to different results. The following, therefore, presents the estimates for all 48 different models that result from these considerations.

7.4.1 Comparison of Estimators and Operationalizations

Figure 7.3 starts with the comparison of the Spatial Lag Y Model as estimator along with the different Halo constellation operationalizations in a coefficient plot. The plot consists of three main columns and three main rows. In the first column, the estimates of the Halo relations with ratios are presented; in the second column, those of the differences; and in the third column, those of the interactions. The rows comprise estimates for the maximum surrounding neighborhood operationalization, the mean operationalization, and the municipality share. Furthermore, each row comprises point estimators for the 1 km^2 direct neighborhoods (unfilled dots) and estimators for the 9 km^2 ones (filled dots). Finally, the point estimators include 95% confidence intervals (thin lines) and 90% confidence intervals (bold lines). In sum, this figure contains 18 different models.

This comparison of the models also yields clear and unambiguous results: no indications for a Halo effect can be found among the Spatial Lag Y Models, even in this range of different operationalizations. As such, it is not the choice of a default operationalization that causes the null-effects of the previous section. In contrast, different operationalizations of the Halo constellations underpin that no effect exists. Moreover, inspecting 10% confidence intervals and directions of point estimates does not even suggest an underlying pattern. All confidence intervals overlap the zero line, and the estimates point in different directions within and between combinations.

By comparing these results with those of the FE model in Figure 7.4, no differences can be reported. All effects are small, they point in different directions, and the confidence intervals overlap the zero line. It should be noted that, according to the FE specification, no municipality level variables can be included in the estimate. Con-



Note: Multiple imputated data from the georeferenced GGSS 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); standardized regression coefficients; all models are controlled for age, gender, education, income, unemployment, home ownership, number of inhabitants in 1 km^2 neighborhood, mean flat size in 1 km^2 neighborhood, size of municipality, eastern or western part of Germany; N = 1,192

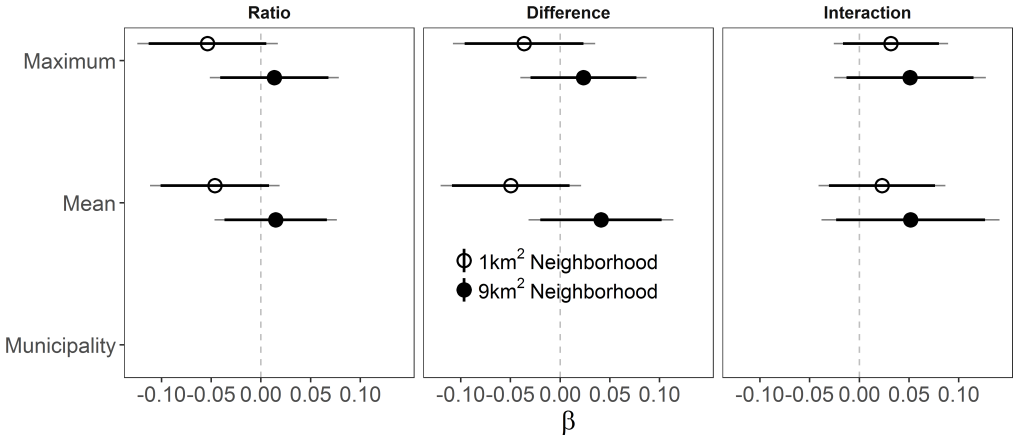
Figure 7.3: Estimates for the Spatial Lag Y Regression Model Across the 3 × 3 Halo Constellations

sequently, no Halo operationalizations with the immigrant share on the municipality level are displayed in the figure. Comparable to the Spatial Lag Y Model, there are no indications for a Halo effect.

Finally, Figure 7.5 shows the estimators for the OLS regression models. Unlike the FE models, they again include estimators for the Halo operationalizations with the immigrant share on the municipality level. Comparable, however, to the previous two figures they show similar results—no Halo effect is found. Although the models with the interaction effect show a slight but statistically insignificant tendency towards the theoretically postulated direction, they should be considered with caution. The OLS models do not control any spatial dependencies, which can lead to overestimated coefficients. Thus, the different OLS models show strong evidence for a non-existing Halo effect.

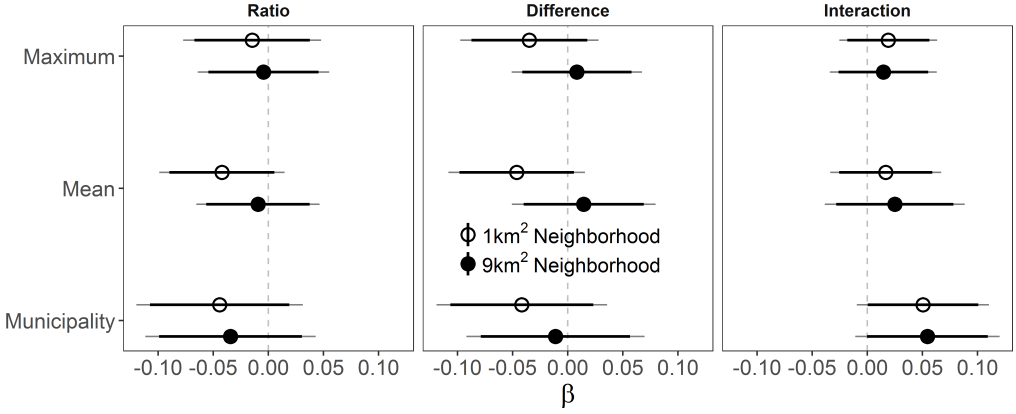
One last reason for a non-existing Halo effect may be the presence of a selection effect. Mainly, in cross-sectional studies, such as this one, and with regards to other hypotheses such as the *White-Flight-Hypothesis* (see, for example, Crowder & South, 2011), the null-effect might occur because xenophobic people already moved away from the Halo neighborhoods. Accordingly, the Halo effect is hidden in the survey data as selection effects blur the real causal effect. In order to navigate this issue, the following section analyzes specific subgroups of people that are at risk to develop xenophobic attitudes but were not able to move away from their neighborhoods. If

there is a Halo effect, the analysis of subgroups should at least partially suggest its existence.



Note: Multiple imputated data from the georeferenced GGSS 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); standardized regression coefficients; all models are controlled for age, gender, education, income, unemployment, home ownership, number of inhabitants in 1 km^2 neighborhood, mean flat size in 1 km^2 neighborhood, size of municipality, eastern or western part of Germany; N = 1,192

Figure 7.4: Estimates for the Municipality Level Fixed Effects Regression Model Across the 3 × 3 Halo Constellations



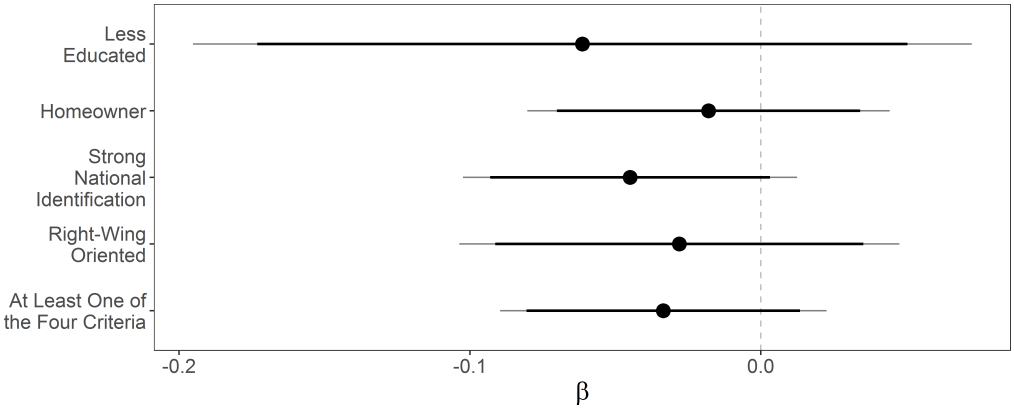
Note: Multiple imputated data from the georeferenced GGSS 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); standardized regression coefficients; all models are controlled for age, gender, education, income, unemployment, home ownership, number of inhabitants in 1 km^2 neighborhood, mean flat size in 1 km^2 neighborhood, size of municipality, eastern or western part of Germany; N = 1,192

Figure 7.5: Estimates for the OLS Regression Model Across the 3 × 3 Halo Constellations

7.4.2 Robustness Checks: Analysis of Subgroups

People who often are considered as being at risk to develop xenophobic attitudes are less educated and unemployed people (Hello et al., 2006), homeowners (Nowak & Sayago-Gomez, 2018), people with strong national identification (Stephan et al., 2009), or people who are right-wing oriented (Bauer, Barberá, Ackermann, & Venetz, 2017). These people would typically move away from neighborhoods in which they feel threatened by immigrants. Thus, it makes sense to conduct subgroup analyses of these people to find indications for a Halo effect within these groups.

Figure 7.6 exhibits such a subgroup analysis[41] in a similar coefficient plot as used before by applying the default operationalization of the Halo constellations and the same default estimator. Again, the results corroborate the null-results from before: no indications for a Halo effect are found. From this perspective, it can be concluded that also among the subgroups of people, who are at risk to develop xenophobic attitudes, Halo effects are not more likely to be observed.



Note: Multiple imputated data from the georeferenced GGSS 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018); standardized regression coefficients; all models are controlled for age, gender, education, income, unemployment, home ownership, number of inhabitants in 1 km^2 neighborhood, mean flat size in 1 km^2 neighborhood, size of municipality, eastern or western part of Germany; $N_{\text{Less Educated}} = 681, N_{\text{Homeowner}} = 696, N_{\text{Strong National Identifikation}} = 869, N_{\text{Right-Wing Oriented}} = 408, N_{\text{At Least One of The Four Criteria}} = 1,152$

Figure 7.6: Estimates for the Default Halo Model Across Five Subgroups That Are Vulnerable to Xenophobia

41 The results for the group of unemployed people were excluded from the figure. Including them in the figure would have stretched the figure's x-axis unnecessarily because the estimates' confidence interval is comparatively large. With a sample size of N=61 and large standard errors of SE=.270 its estimate of β=-.122 is not reliable.

7.5 Discussion

Social interaction between people of different origins reduces prejudices, but it can also lead to feelings of threat and xenophobia when it remains casual and superficial. After decades of research, it is still not clear how the amount of immigrants in a specific neighborhood relates to this problem, whether it increases or decreases prejudices and xenophobic attitudes. The Halo Hypothesis offers a solution by emphasizing the relationship between direct and surrounding neighborhoods: ethnically diverse neighborhoods increase xenophobia when they border ethnically homogeneous neighborhoods. From the results of this application, however, it must be reported that there is no Halo effect in German neighborhoods.

To substantiate this finding, different operationalizations of neighborhoods and estimators to adjust potential biases are used. Varying neighborhoods' geographic sizes, their relation to each other, and adjusting dependence between different survey respondents and between different municipalities all lead to null-results. These findings are surprising, considering the successful application of the Halo Hypothesis in the research of others (Martig & Bernauer, 2016; Rydgren & Ruth, 2013). Moreover, the estimation of the null-results is rather precise, which makes similar findings in the future more likely. What remains are methodological considerations and also some substantive reasons that might explain these findings.

7.5.1 Methodological Considerations

First, one of the most important methodological considerations concerns the measurement of the dependent variable xenophobia. This variable is created by inserting five items on attitudes towards immigrants in an PCA and by extracting the factor scores for the first component. Other research on the Halo effect, however, used different dependent variables, such as election results for right-wing parties. Differences in the results could exist because of structural differences between xenophobic attitudes and the voting for right-wing parties. At the same time, variables in this application such as education and unemployment show similar results in studies of political behavior, and there are no theoretical reasons which explain why the Halo effect appears in the case of political behavior and not in those of political attitudes. Consequently, it is questionable whether disparities between political attitudes and political behavior explain why a Halo effect does not occur in the former but the latter.

Second, a more pressing concern is the operationalization of the Halo constellations. They are based on immigrant rates in specific spatial units of 1 km^2 of the Ger-

man Census 2011 but not on information about people with specific migration backgrounds. The group of people with migration background may be larger than the group of immigrants, and also considering specific subgroups may have been more instructive. For example, studies concerned with ethnic heterogeneity in neighborhoods often investigated populations from non-Western origins (see, for example, Morales & Echazarra, 2013). Although the amount of immigrants in a specific neighborhood also correlates highly with the number of people with migration background (Koopmans & Veit, 2014), following studies could try to disentangle both population types with more detailed data that might be available in the future.

Third, despite the analysis of subgroups, a selection effect cannot be ruled out entirely. In order to investigate this issue, panel data are required with information on residential moving. The argument for the White-Flight-Hypothesis is rather convincing, and the subgroups in this study were small. Even in these groups, however, the null-results have been estimated precisely. Nevertheless, more sophisticated modeling techniques which make use of panel structures (e.g., as in Lancee & Schaeffer, 2015) would give more information about why this is the case.

Fourth, this analysis cannot control for friendship or contact between Germans and immigrants. Accordingly, controlling social interactions between people was not possible in this application, but this variable would have been essential to examine the exact mechanisms of interactions. At the same time, by varying the geographic scale of the Halo constellations, a suppression of this variable can be ruled out. If it was acting as a suppressor, mainly in the case of positive contact between in- and out-group members, the estimates in this analysis would have been overestimated because theory implies that positive contact decreases xenophobia.

7.5.2 Segregation Structure of Germany

The results of this empirical application imply that no Halo effect exists in Germany. Although some methodological considerations suggest that the Halo effect should be pursued further with different data, some substantive reasons for this null-effect may be relevant as well. As all other studies on the Halo effect were conducted in other societies than the German one, it can be possible that the null-findings relate to the different social structures of Germany and other countries.

Germany has a comparatively low level of ethnic segregation compared to other countries (Koopmans, 2010; Schönwälder & Söhn, 2009). As a consequence, people do not perceive neighborhood boundaries so strongly. The assumption of the Intergroup Threat Theory that people may fear competition with foreigners from border-

ing neighborhoods in the economic and cultural context does not apply—hence, there is no public discourse about this topic in Germany. This circumstance does not rule out the possibility that some groups feel threatened by foreigners, but this fear is not based on distinct neighborhood boundaries. There is simply no social frame about competing neighborhoods. While Halo constellations empirically exist, they may not be present in the perception of people in these neighborhoods.

7.5.3 Conclusion

In sum, the results of this application suggest how sensitive social science theories are to differences in societal contexts. While to a certain degree the null-findings may also be caused by methodological considerations that require further investigation, the substantial argument proposes that they are caused by the residential structure of Germany. Social science theories always deal with social contexts, but they also differ in their predictions depending on their application in specific social settings. The georeferenced survey data of this application provide in-depth insights into the state of spatial integration of such theories.

8 Application III: Social Inequalities of Environmental Hazards Exposure

8.1 Research Question

Of all the stressors intruding people's lives, some of the most prominent are environmental hazards. Although environmental hazards that derive from, e.g., industry emissions decreased over the last decades in Germany (Federal Agency for the Environment Germany, 2015, 60), those originating from land use, such as soil sealing or lacking recreational green spaces, are continually growing. These hazards—environmental bads or lacking environmental goods—do not only involve environmental issues but also give rise to societal questions relating to residential segregation and inequality, mainly because they affect people's well-being (Gidlöf-Gunnarsson & Öhrström, 2007), stress processing (Thompson et al., 2012) as well as their physical (Schulz, Romppel, & Grande, 2016) and mental health (Guite et al., 2006). For these reasons, researchers have been investigating the social and spatial inequalities of local environmental goods and bads already for a long time. They generally found that environmental hazards apparently affect more people of low socio-economic status or ethnic minorities (Braubach & Fairburn, 2010; Rüttenauer, 2018; Wolch, Byrne, & Newell, 2014). Thus, given a continually growing land use in Germany, environmental inequalities remain a subject of an ongoing societal and scientific debate.

However, there are two reasons why previous findings on environmental inequalities—most of them gathered within the US context—cannot directly be transferred to the German societal context. First, the residential segregation structures of Germany and other countries, such as the US, the UK or the Netherlands, differ. Ethnic residential segregation in Germany is not as strongly concentrated as in other societies in which large urban neighborhoods of unique ethnic backgrounds exist (Schönwälder & Söhn, 2009). Second, what is known about residential segregation in Germany is little because of missing data (Goebel & Hoppe, 2015), especially with regards to environmental hazards. Accordingly, findings on environmental inequalities in Germany are rare (Best & Rüttenauer, 2018; Braubach & Fairburn, 2010) or lack more in-depth investigation, primarily using objective data on a smaller geographic scale. This application aims to close this gap by examining the mechanisms of environmental inequalities in Germany in more detail using survey data combined with small-scale spatial data on land use in Germany.

8.1.1 A Matter of Social and Ethnic Inequality?

Past research discussed environmental inequalities with regards to different theoretical considerations of residential segregation. Broadly, these considerations are either arguments stressing the socio-economic mechanisms of residential segregation or arguments underlining the social processes behind residential segregation. In Germany, findings are supporting both of these arguments (Best & Rüttenauer, 2018). This section discusses the explanations of segregational and associated environmental inequalities and determines their use to the German societal context.

The *Socio-Economic Inequality Thesis* states that property values and rents determine which social groups move to which location (Downey et al., 2016). Accordingly, members of, e.g., low-income groups move to neighborhoods where property values are small and rents are low. In contrast, members of high-income groups can afford to move to neighborhoods where property values and rents are high. The higher the socio-economic inequalities in a society, the more pronounced are residential patterns.

The reason for low property values and rents in certain neighborhoods is, among others, the prevalence of environmental hazards within or near these neighborhoods. Examples are residential areas exposed to industrial hazards (Crowder & Downey, 2010) or traffic noise (Bocquier et al., 2013), as well as areas characterized by missing access to green spaces (Gidlöf-Gunnarsson & Öhrström, 2007; Guite et al., 2006). Neighborhoods with high property values and rents are less polluted, quieter and have higher amounts of green spaces. Thus, the Socio-Economic Inequality Thesis predicts that members of low-income groups are more frequently exposed to environmental hazards than members of high-income groups.

Similarly, the *Ethnic Income Inequality Thesis* states that ethnic inequalities relating to residential segregation patterns occur because members of ethnic minorities often have a lower income (Crowder & Downey, 2010). In consequence, members of ethnic minorities move to neighborhoods where property values are small and rents are low; members of the ethnic majority move to neighborhoods where property values and rents are high because they can afford these housings. Controlling income should reduce differences between different ethnic status groups.

In contrast to the Socio-Economic Inequality Thesis and the Ethnic Income Inequality Thesis, competing theories include the concept of assimilation. Rather than stressing property values and rents, the *Spatial Assimilation Hypothesis* says that moving also involves a process of seeking for matching neighborhoods (Crowder & Downey, 2010, 5). Accordingly, members of specific social groups—low or high income, ethnic group A or ethnic group B—make choices for neighborhoods where residents share similar

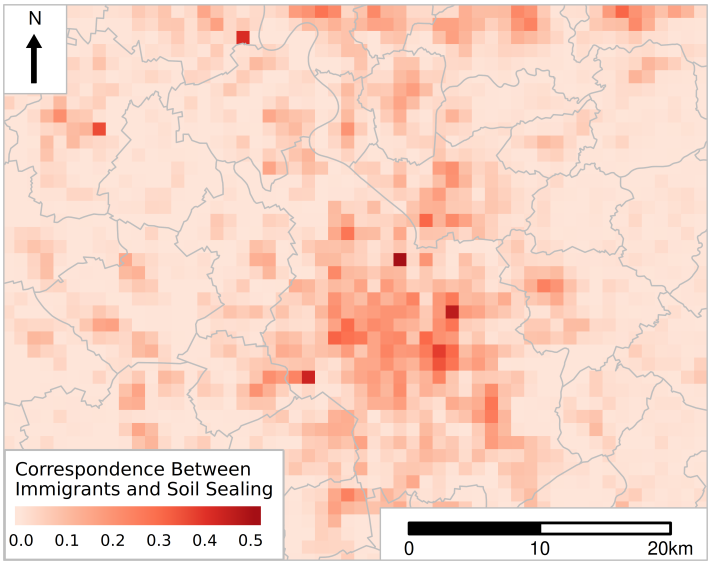
characteristics as themselves. Lersch (2013), for example, showed for the German context that people with a Turkish migration history are more likely to choose neighborhoods which match their migrant background. Unlike other social groups, this idea is more important to them than choosing a different neighborhood with higher quality. Controlling income may not reduce differences between different migrant groups and their residential segregation patterns, which, however, cannot be tested here because it requires longitudinal data.

Finally, another prominent explanation for inequalities of exposure to environmental hazards, primarily relating to ethnicity, is that moving is not necessarily a product of voluntarism. In their pioneering work for the Swiss societal context, Diekmann and Meyer (2010) found that, after controlling all relevant socio-economic factors, people with foreign citizenships are still more probable to be exposed to air pollution or road traffic noise than other people. These differences may be caused by discrimination in housing markets (Schönwälder & Söhn, 2009, 1452; Auspurg, Hinz, & Schmid, 2017) which systematically pushes members of ethnic minorities to neighborhoods exposed to environmental hazards, independent of their socio-economic status (Lersch, 2013). Hence, segregation patterns and environmental inequalities are a multidimensional phenomenon consisting of both person-level as well as contextual factors.

Which of these mechanisms and competing hypotheses applies to potential environmental inequalities in Germany? Socio-economic factors can shape residential segregation patterns, and so can factors of assimilation and discrimination. Figure 8.1 illustrates that conventional analyses cannot help to detect the exact mechanisms. On a map of the city of Cologne, it displays measures of correspondence between immigrant rates and the density of soil sealing in 1 km^2 neighborhoods. Accordingly, differences in exposure to environmental hazards exist at least between neighborhoods with different immigrant rates, but from these data, it cannot be deduced which proposed mechanism is in place.

This empirical application tests the competing hypotheses of the Socio-Economic and Ethnic Income Inequality as well as the Ethnic Discrimination Thesis to explain these differences of environmental hazard exposure. While Best and Rüttenauer (2018) demonstrated that factors such as income might not play a role in the link between migration status and subjectively perceived exposure to environmental hazards, evidence with objective measures is still lacking. Moreover, despite decades of immigration history in Germany, evidence on residential segregation is still rare (Goebel & Hoppe, 2015, 23f). Thus, another aim of this application is to contribute to this research. Before going any further, what is known about residential segregation in Germany is summarized, and its implications for this application's research ques-

tion is discussed.



Data Sources: Leibniz Institute of Ecological Urban and Regional Development (2018) and Federal Agency for Cartography and Geodesy (2018); the values of both data sources were normalized to a scale between 0 and 1, and multiplied

Figure 8.1: Correspondence Between the Amount of Immigrants and Soil Sealing Density in 1 km^2 Neighborhoods of the City of Cologne and Surrounding Municipalities

8.1.2 Residential Segregation Structure in Germany

The residential segregation structure of immigrant groups in Germany differs in comparison to other countries. While segregational differences between immigrants and native Germans exist, ethnic concentration in urban neighborhoods is not as pronounced as, e.g., in the Netherlands or the UK (Schönwälder & Söhn, 2009). At the same time, neighborhoods with higher rates of immigrants are more frequently affected by higher rates of unemployment and dependence on welfare. Thus, although residential segregation is not as concentrated in Germany as in other countries, inhabitants of neighborhoods with comparatively high immigrant rates are at higher risk to experience inequalities in the form of social deprivation (Goebel & Hoppe, 2015). Consequently, among the groups exposed to environmental hazards, it is expected to find a higher share of people with migration background. This hypothesis derives from the expectation of higher environmental hazard rates in low-income neighborhoods (Zwickl et al., 2014).

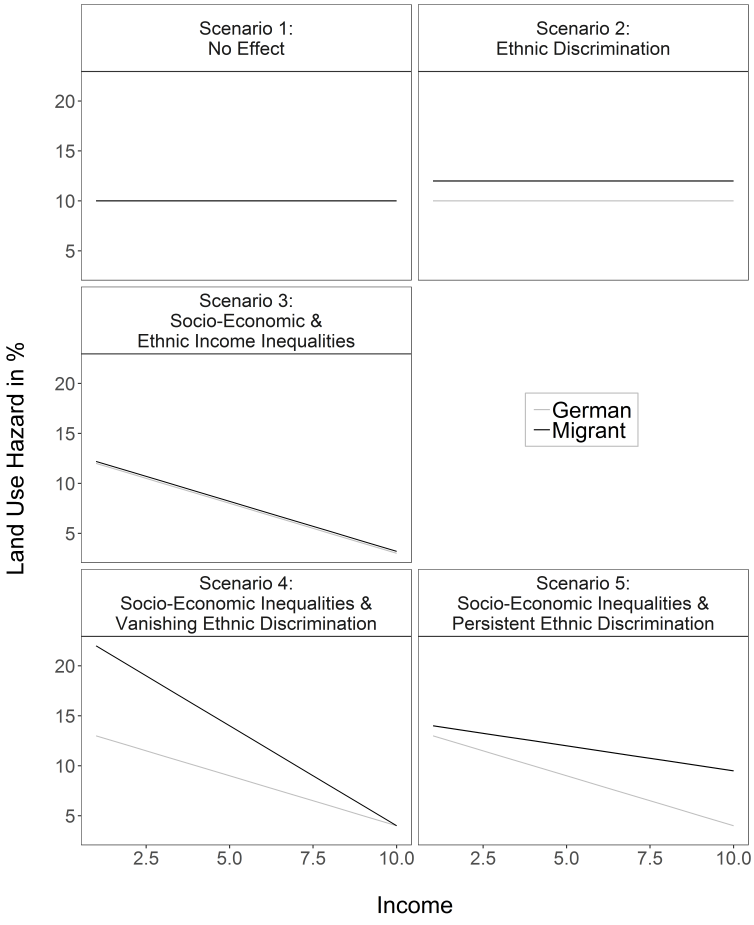


Figure 8.2: Possible Outcomes of Land Use Hazards as a Function of Income Depending on the Competing Hypotheses

The partly competing theoretical considerations of this empirical application result in five possible scenarios shown in Figure 8.2. Scenario 1 applies when no relationship between migrant status, income, and environmental hazard exposure exists—corresponding to a null-result. Scenario 2 differs from scenario 1 in the detail that ethnic inequality exists, but income does not reduce the exposure for both groups (Ethnic Discrimination). Scenario 3 again differs in such a way that all inequalities between groups can be explained by income. Thus, income reduces environmental hazards for both groups (Socio-Economic & Ethnic Income Inequalities). Both scenarios 4 and 5 weaken the assumptions of scenario 3 either by a discrimination of migrants in low-income groups (Socio-Economic Inequalities & Vanishing Ethnic Discrimination) or by a still existing discrimination of migrants especially in high-income

groups (Socio-Economic Inequalities & Persistent Ethnic Discrimination).

Each of these scenarios emphasizes the role of income. First, if the Socio-Economic and Ethnic Income Inequality Thesis hold, higher income decreases the risk of exposure to land use hazards. Second, if the Ethnic Discrimination Thesis holds, income does not play a significant role at all. Figure 8.2 displays these competing considerations: if income does not play a role in land hazard exposure, the exposure remains static and does not change, neither for German nor for the migrant group. If, however, income plays a major role, the exposure to environmental hazards decreases in both groups.

8.1.3 Small-Scale Spatial Data Studies and Environmental Inequalities Research

Generally, when studying the context of social behavior, a small scale perspective on this context is preferable to detect people's patterns of social behavior. Studies in an extensive range of different disciplines, however, often only can draw on aggregated data of comparatively large administrative units, such as municipalities or regions. The problem with these data is that, for example, ecological fallacies are more probable when statistical relationships found on the aggregated level are transferred to the individual level (Bluemke et al., 2017; see also Chapter 3.2.3). Thus, using data containing measures of the direct living environment of people prevents ecological fallacies and can lead to more valid model estimations (Sluiter et al., 2015).

In particular, research on environmental inequalities benefits from such data. Crowder and Downey (2010) demonstrated how such data support the careful reassessment of previous ambivalent findings and the tracing of people's moving behavior. Likewise, Downey et al. (2016) showed how environmental inequalities between specific demographic groups, in this case, two-parents and single-parent families, can be better understood with small-scale data. This empirical application of the Chapter at hand also uses person-level survey data spatially linked to small-scale geospatial data on a 100 meters $\times$ 100 meters raster grid level. Moreover, this level is varied by applying different sizes of geographic buffers (see Chapter 5.2.2)

Other studies that used land use data have motivated the choice of these data for this application. It has been shown, for example, that people living close to industry or trade facilities report poorer health (Marques & Lima, 2011). People who live near to green areas, in turn, profit from lower levels of stress hormones (Thompson et al., 2012). Both, environmental goods and environmental bads, thus can be found among information about the use of land. As shown in Chapter 2.4.2, soil sealing qualifies as

a general indicator for environmental hazards because it subsumes lacking environmental goods but also environmental bads. Accordingly, it is assumed that an unequal distribution of soil sealing among distinct social and ethnic groups is a direct indicator of environmental inequalities among these groups. This application aims to find out whether this also applies to the German societal context.

8.2 Data and Measures

8.2.1 Sample: GESIS Panel and IOER Monitor Data

The data for this empirical application stem from the combined data of the georeferenced GESIS Panel (see Chapter 2.3.2) and geospatial data about soil sealing from the IOER Monitor (see Chapter 2.4.2). These data were linked by location and by using buffers to create varying geographic sizes of potential environmental hazard exposures. Overall, the sample includes data from 3,852 survey respondents.

The following section presents in detail how the variables were calculated. Comparable to the other empirical applications in Chapter 6 and 7, it starts with an introduction to the geospatial data measures, before the survey data measures are introduced. A table of descriptives of all measures is displayed in Table 8.1 at the end of this section.

8.2.2 Geospatial Data Measures

Geospatial Data Preparations

Similar to the German Census 2011 data that were used in the empirical application of Chapter 7, land use data from the IOER monitor are already available in a harmonized format (see also Chapter 2.4.2). Moreover, the data can directly be accessed from web services which the IOER provides, and are already prepared in spatial data formats (the German Census 2011 data are available as CSV files that first have to be prepared by the users). No matter which attributes of the IOER monitor are to be used, they do not need to be converted in a raster data format.

To request these data from the website of the IOER, however, tools to access the Application Programming Interface (API) are needed.[42] For this purpose, either graph-

42 A current research project, funded by the German Research Foundation, facilitates these requirements for social science researchers in Germany. The project consortium consisting

ical tools from GIS software such as *QGIS* can be used to connect to the web service and download the spatial data layers; or programming languages, such as *Python*, can be deployed, which then also create reproducible code. Generally, for both methods, users need to know the Uniform Resource Locator (URL) of the service, the name of the indicator, the year, and the spatial resolution. In this application, the URL is gathered from the IOER monitor website (www.maps.ioer.de) and the indicator of soil sealing is chosen for the year 2014 in a geographic resolution of 100 meters $\times$ 100 meters.

Spatial Buffers of Varying Sizes

Land use is a direct consequence of people interacting with their environment. Settlement in areas leads to the building of houses and roads, and depending on supply and demand, new buildings and roads are added. Most of the settlements are a cause of historical development, and disentangling all processes of the human interaction with the environment is difficult. However, also in the context of environmental hazards it is obvious that this interaction is two-dimensional: land use is not only a consequence of human behavior but also influences the life of people because they are affected by lacking green spaces (Thompson et al., 2012) and existing "grey areas" of concrete and walls in cities (Marques & Lima, 2011). Comparable to air pollution and road traffic noise, land use is an environmental hazard.

What remains challenging is to operationalize environmental hazards from available land use data. These data were collected to capture detailed information on land use for large extents of a geographic area, but generally, they do not correspond to geospatial units such as neighborhoods or other administrative borders. Researchers have to question what the socially relevant context (see Chapter 2.6) of land use might be. While in some research applications smaller geographic areas may be preferable (Nonnenmacher, 2013), it is not yet decided whether this applies to land use as well. Choosing the right geographic size of environmental influences of land use, at least to some extent, is an empirical question.

Because of these considerations, in this empirical application, ego-hoods are created by using spatial buffers for soil sealing (see Chapter 3.2.1 and 5.2.2). Besides the

---

of the GESIS - Leibniz Institute for the Social Sciences, the Leibniz Institute for Ecological and Urban Development, the Karlsruhe Institute for Technology and the Socio-economic Panel build an easy to access technological infrastructure to combine social science survey data and geospatial data without the preknowledge of geospatial data and GIS. Details can be found on the project's website: www.sora-projekt.de.

use of soil sealing in 100 meters × 100 meters neighborhoods, subsequently, buffers of the sizes of 500 meters, 1000 meters, and 2000 meters are created. The idea is not necessarily to focus on the size of the buffer with the best model fit (Spielman & Yoo, 2009, 1102) but to analyze how predictions from statistical models change with different geographic sizes of ego-hoods.

8.2.3 Survey Measures

Income

To test the hypotheses on social inequalities and income, this application uses an income variable of the GESIS Panel that relies on categorized values of the household income in EUR: 0 - 299 = 1; 300 - 499 = 2; 500 - 699 = 3; 700 - 899 = 4; 900 - 1099 = 5; 1100 - 1299 = 6; 1300 - 1499 = 7; 1500 - 1699 = 8; 1700 - 1999 = 9; 2000 - 2299 = 10; 2300 - 2599 = 11; 2600 - 3199 = 12; 3200 - 3999 = 13; 4000 - 4999 = 14; 5000 - 5999 = 15; 6000 - 9999 = 16; 10000 and above = 17. An advantage of using these categories is that fewer observations are missing in comparison to more detailed income measures. Furthermore, the distribution of the variable is normal. Thus, no transformations of the original scale are necessary for later analyses. For convenience, this measure of household income is referred to as income in the following sections.

Migration Background

The migration status of the respondents is based on the birth country of their fathers. Using the birth country of parents as operationalization guarantees to also include information about people who are German citizens but still have a migration history in the second generation. The dichotomous variable of all respondents whose fathers were not born in Germany is coded as 1, whereas those with a German father are coded as 0.

Sociodemographic Controls

The analysis includes two more sociodemographic variables: the age of the respondents and their declared gender. Some studies suggest that older people tend to live in more rural neighborhoods (Shepherd et al., 2013, 1289). As all used indicators also confound with urbanization, the analysis controls for the age of the respondents. With regards to the gender of the respondents, a gender-based tendency to move to neigh-

borhoods with higher rates of land use hazards is not expected. Previous research, however, showed that among single parent families, women are at higher risk to be exposed to environmental hazards (Downey et al., 2016). For this reason, the gender of the respondents is controlled as well.

Table 8.1: Descriptive Statistics and Overview of all Variables of the Analysis (Pairwise Deletion)

	Mean / %	SD	Minimum	Maximum
Dependent Variables				
Soil Sealing (100m × 100m)*	57.06	26.70		
Soil Sealing (500m Buffers)*	39.44	23.05		
Soil Sealing (1000m Buffers)*	30.87	21.87		
Soil Sealing (2000m Buffers)*	23.24	19.47		
Independent Variables				
Income	11.70	2.84	1	17
Migration Groups:				
German	90.02			
Migrant	9.98			
Individual Controls				
Age	46.15	14.08	18	73
Gender (Female)	52.39			
Education:				
Low	21.35			
Medium	34.98			
High	43.67			
Homeownership	53.88			
Household Size	2.65	1.24	1	14
Inter-Individual Controls				
Number of Inhabitants*	212011.75	576284.72		
Number of Observations	3852			

Data Source: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); * some values removed due to data protection

Other Contextual Controls

Also, the analysis adjusts its estimates for several other variables that influence moving behavior. These controls include the highest achieved education of the respon-

dents (low, medium, and high) as well as the number of people living in the same household as it is assumed that the size of the household also determines the opportunities to move to specific neighborhoods. Furthermore, a dichotomous indicator for whether the respondents live in their household for rent or whether they are homeowners is incorporated because this may influence how people have chosen where they want to live. Finally, the analysis contains a control variable for the number of inhabitants in the respondents' municipality. The latter will also be used as an additional predictor in some robustness checks later.

8.3 Analysis Strategy: Linear Prediction Models

Can the analysis find patterns of environmental inequalities between the native population and migrants in Germany? Moreover, if so, can these inequalities be explained as a function of income? The analysis starts with a baseline model that yields the results for the general exposure to land use hazards by soil sealing among the population (by location, buffers of the size of 500, 1000, and 2000 meters). Subsequently, the analysis continues with a series of linear regression models for soil sealing that include interactions of migrant status and income. The results of this analysis are presented as predicted values in separate figures because their interaction effects may be hard to interpret. The same holds for the final robustness check that incorporates interactions between migrant status, income and the inhabitant sizes of municipalities.

In order to prevent the prediction of values over 100% or below 0% for the dependent variable soil sealing, a logit transformation is used before the estimation:

$$y_{logit} = ln\frac{y}{1-y} \tag{8.1}$$

This transformation bounds the predicted values of the dependent variables to an interval between -3.663562 and 3.663562. After re-transforming the results of the predictions to the original scale of percentages, their values lie exactly between 0% and 100%—a better interpretable value of predicted soil sealing. The re-transformation is applied as follows:

$$\hat{y} = \frac{e^{\hat{y}_{logit}}}{1 + e^{\hat{y}_{logit}}} \tag{8.2}$$

Moreover, all linear regressions are estimated using cluster-robust standard errors. Different residential policies between municipalities may produce clustering and dependencies between respondents within municipalities. These dependencies result in underestimated standard errors that increase the risk of conducting statistical er-

rors of type I (Abadie et al., 2017). In order to navigate these issues, the analysis is adjusted for clustering across the municipalities (N = 373).

Lastly, this multivariate analysis is based on 100 multiple imputed datasets by using sequential regression imputation (Raghunathan, 2016, 67f) and all variables on the respondents' level as input data, which is advertised as providing efficient estimates for data with large proportions of missing data (Graham, Christian, Kiecolt-Glaser, & Ader, 2007). Alternative estimations with list-wise deletions corroborate this approach and yield similar results but with less statistical precision (see Table C.1 of the Appendix).

8.4 Results

8.4.1 Baseline Model of General Soil Sealing Exposure

As outlined above, the analysis starts with a general inspection of the soil sealing exposure. Moreover, this analysis also provides estimates for all other predictors such as age or homeownership and how they relate to the dependent variable soil sealing. In sum, this results in a series of four linear regression models with cluster-robust standard errors on the dependent variables varying by geographic size.

Table 8.2 displays the results of these linear regression models. First of all, the model estimates differ in effect sizes as well as the level of statistical significance among the covariates. Models with soil sealing measured by increasing geographic sizes yield more pronounced results than the other models. However, in all four models, a positive coefficient for the group of migrant people can be observed. In comparison to the group of Germans, soil sealing hazards affect migrant people more. These results show that ethnic inequality with regards to soil sealing exists.

Also, the estimates for the other predictors in the models are instructive. While economic factors such as homeownership or income decrease soil sealing hazards—although the latter is not statistically significant—, this may not hold for education. In particular, groups with high education, in comparison to those with low education, live in areas with higher levels of soil sealing. Thus, while being associated with economic wealth, education cannot always be transferred into better housings. Accordingly, it is of particular interest how income moderates the found evidence for ethnic inequalities, albeit its effect is low in the baseline model.

Table 8.2: Standardized Regression Coefficients for the Baseline Model Between Increasing Geographic Sizes of Soil Sealing (N = 3,852; Clustered Standard Errors)

	100m × 100m			500m Buffer			1000m Buffer			2000m Buffer		
	β	SE	CI	β	SE	CI	β	SE	CI	β	SE	CI
Intercept	1.077***	.162	[.759, 1.394]	-.207†	.125	[-.452, .038]	-.950***	.147	[-1.239, -.662]	-1.551***	.150	[-1.845, -1.257]
Age	-.005*	.002	[-.008, -.001]	.000	.001	[-.003, .003]	.002	.002	[-.001, .005]	.002	.002	[-.001, .005]
Gender (Female)	.079†	.045	[-.010, .167]	.066†	.034	[.000, .132]	.088*	.041	[.009, .168]	.090*	.038	[.015, .164]
Education												
Low (Ref.)												
Medium	-.044	.070	[-.182, .094]	-.007	.057	[-.118, .104]	-.003	.070	[-.141, .135]	.008	.072	[-.134, .150]
High	-.065	.067	[-.197, .068]	.139**	.054	[.034, .244]	.276***	.069	[.141, .410]	.337***	.070	[.200, .474]
Homeownership	-.426***	.066	[-.556, -.297]	-.453***	.049	[-.550, -.357]	-.544***	.057	[-.655, -.432]	-.545***	.057	[-.657, -.434]
Household Size	-.089***	.022	[-.133, -.045]	-.089***	.016	[-.120, -.057]	-.104***	.019	[-.142, -.066]	-.097***	.019	[-.134, -.060]
Inhabitants	.000**	.000	[.000, .000]	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]
Income	-.015	.010	[-.035, .006]	-.006	.008	[-.022, .010]	-.004	.010	[-.023, .016]	-.003	.010	[-.023, .017]
Migrant	.183*	.081	[.025, .341]	.335***	.064	[.209, .461]	.465***	.074	[.321, .609]	.417***	.075	[.269, .565]

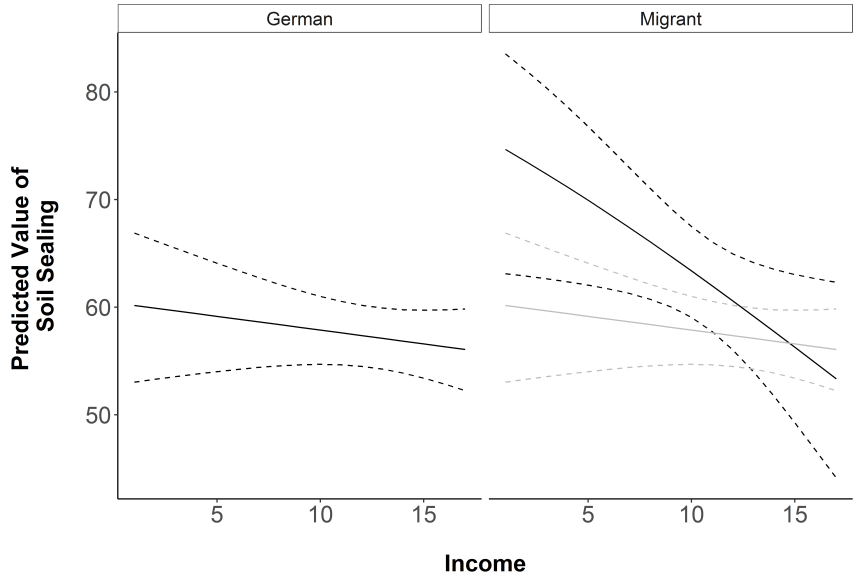
† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017)

8.4.2 Soil Sealing Exposure as a Function of Income

How does income relate to the differential soil sealing exposure of migrant people? By predicting each geographic size of soil sealing with varying levels of income and holding all other predictors constant at their mean value, the following analysis reveals patterns of possible relationships. The results of these predictions can directly be compared with the five scenarios shown in Figure 8.2 and are shown in four separate figures—one for each of the varying geographic sizes of the neighborhoods for soil sealing hazards.

Figure 8.3 starts with the prediction results for the analysis of soil sealing on a geographic scale of 100 meters × 100 meters neighborhoods. The figure consists of separate plots for the group of German people and the migrant group. 95% confidence intervals display the uncertainty in the predictions. In order to facilitate finding differences between the group of German people and migrant people, the predictions for the former are also displayed within the subplot of the migrant group.

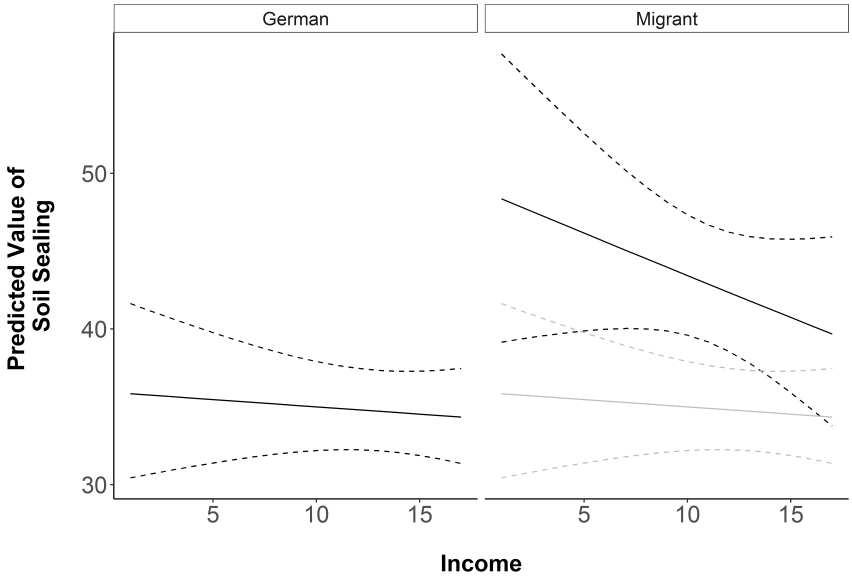


Note: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

Figure 8.3: Predicted Values for Soil Sealing as a Function of Income for German and Migrant People in 100m × 100m Neighborhoods

Generally, for both groups, the figure displays a decreasing slope for income. People

with increasing income live in neighborhoods that are less exposed to soil sealing. While this slope is moderate for the German group, for the group of migrants it is rather steep. At the same time, migrant people start with a higher initial level of soil sealing in the low-income group and approximate the level of the German group in the high-income group. These difference, however, lack statistical precision and are not statistically significant on a level of $p \leq .05$ because both groups' confidence intervals overlap. As 100 meters $\times$ 100 meters neighborhoods are comparatively small it remains interesting how income relates to soil sealing in larger ego-hoods.



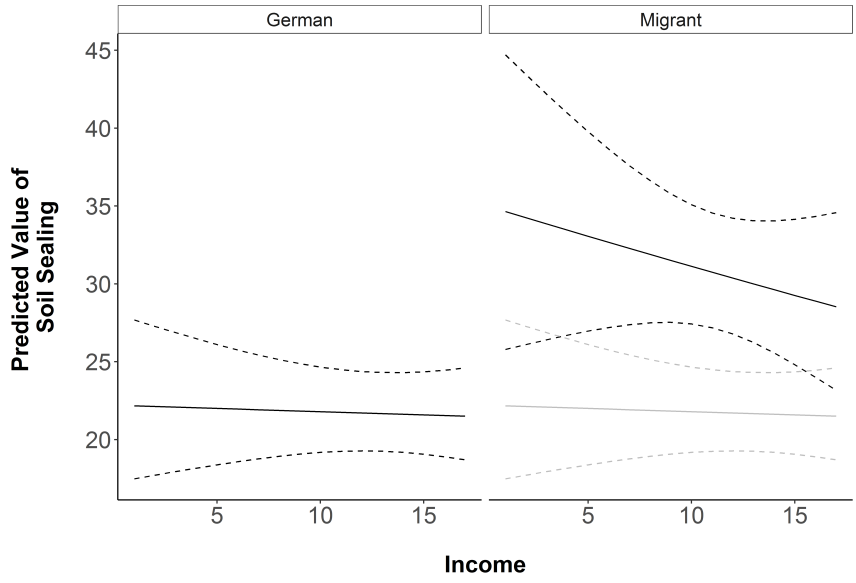
Note: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

Figure 8.4: Predicted Values for Soil Sealing as a Function of Income for German and Migrant People in Ego-Hoods with a 500 Meters Buffer

Figure 8.4 shows the results for the analysis of soil sealing in ego-hoods with a 500 meters buffer. What immediately stands out is that for both groups the initial level of sealing soils is much lower than in the 100 meters $\times$ 100 meters neighborhoods. Also, for both groups soil sealing exposure decreases with increasing income, although the slope is again steeper for migrants than for Germans. In contrast to the results of the geographic scale of 100 meters $\times$ 100 meters, in the medium income class, the figure now shows differences between the German and the migrant group—indicated by the gap between the lower bound of the migrant group's confidence interval and the

upper bound of the German group's one. Even in income classes above the median of 12, migrant people are more exposed to soil sealing hazards in ego-hoods with a 500 meters buffer.

This result is corroborated by Figure 8.5 and 8.6 which comprise the estimates for ego-hoods with increasing buffers of 1000 meters and those of 2000 meters. Mainly in the case of ego-hoods with a 1000 meters buffer, the gap between the German and migrant group is considerably high, indicated by the non-overlapping confidence intervals. The difference between the confidence intervals gets smaller in ego-hoods with a 2000 meters buffer, but so does also the initial level of exposure to soil sealing. Generally, in both figures, the initial level of exposure drops significantly. This finding may suggest that increasing the buffer sizes of neighborhoods even more, would not make any sense. The results do not gain any more precision.



Note: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

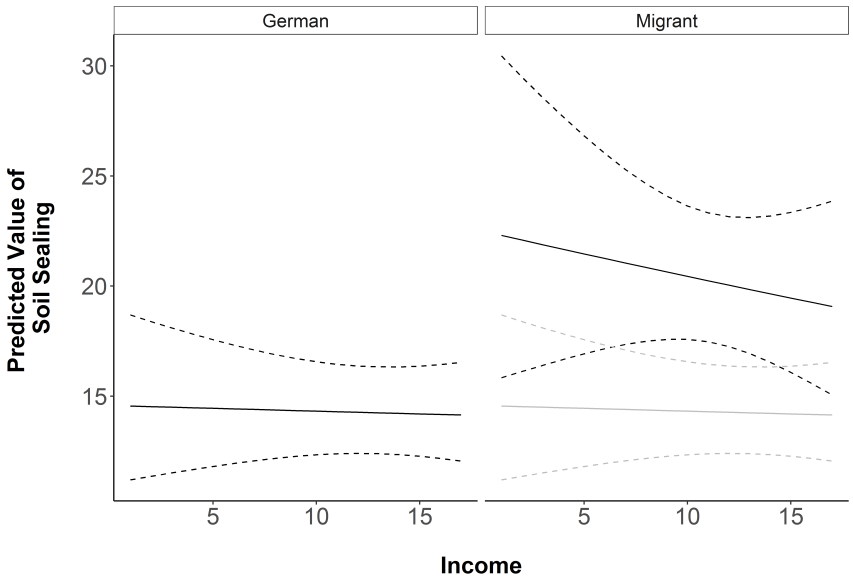
Figure 8.5: Predicted Values for soil sealing as a Function of Income for German and Migrant People in Ego-Hoods with a 1000 Meters Buffer

The series of analyses proves the following:

1. soil sealing hazards affect people in low-income groups more than in high-income groups. Thus, income decreases exposure to land use hazards.

2. People from the German majority group are generally better off with regards to these soil sealing hazards. Even when their general decrease in exposure is lower, their initial level of exposure is considerably smaller than for the migrant group.

3. The effects of ego-hoods with buffers of $\geq$ 500 meters are the most pronounced ones as they show statistically significant differences between the German and the migrant group, but they start to decrease again with buffers of 2000 meters.



Note: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); imputed, predicted and combined using Rubin's Rule; 95% confidence intervals based on cluster robust standard errors; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 3,852

Figure 8.6: Predicted Values for Soil Sealing as a Function of Income for German and Migrant People in Ego-Hoods with a 2000 Meters Buffer

To summarize these findings with regards to their theoretical interpretation, a combination of the Socio-Economic and Ethnic Income Inequalities Thesis, as well as the Ethnic Discrimination Thesis, may be the appropriate explanation for the findings 1 and 2. The results show that differences in general exposure decrease with increasing income between German and migrant people. At the same time, the initial levels of land use hazards exposure are higher for the migrant group, and even for people with high income within the migrant group, they may still exist. Thus, even if income explains some of the inequalities between German and migrant people, discrimination, e.g., on the housing market, as indicated by others (Auspurg et al., 2017; Diekmann &

Meyer, 2010), could still be in place.

Another explanation concerns spatial assimilation or spatial segregation. Some unobserved subgroups within the migrant group may mainly live in neighborhoods that are exposed to land use hazards, as they happen to live in municipalities with higher population densities. In Germany, population density on an aggregated level correlates with immigrant rates.[43] The estimated models are controlled for inhabitant sizes on the municipality level, but they are kept constant at their mean level of 212,011.70. Inhabitant sizes might cover patterns of residential segregation in the models when they are not varied. For example, the relationship between income and soil sealing exposure could differ between municipalities with higher and lower inhabitant sizes. Likewise, the interaction between this relationship and the migrant group status could differ as well. The following final analysis step inspects this suspicion of differences among different inhabitant sizes.

8.4.3 Robustness Check: The Role of Municipalities' Inhabitant Sizes

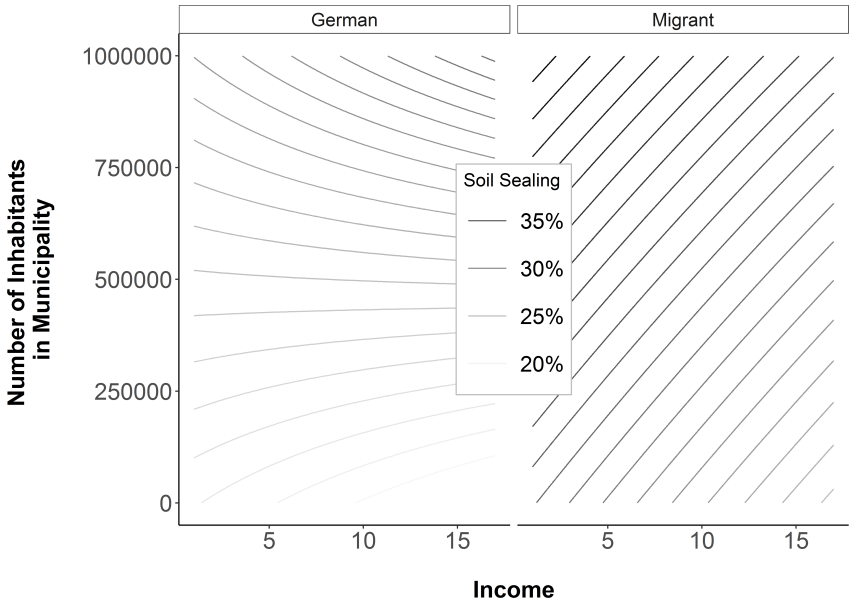
Analyzing the role of municipalities' inhabitant sizes in the relationship between income and soil sealing hazards among German and migrant people leads to multivariate interaction effects. Depending on the manifestation of income and inhabitant sizes, the predicted value of soil sealing changes. Formally, this results in three-dimensional regression planes for each group. Interpreting such regression planes requires different viewing angles that must be alternated. An alternative is using contour plots which project the predicted values of a third variable onto a two-dimensional space. Contour plots are, therefore, an appropriate option to present the following multivariate interaction effects.

Figure 8.7 shows such a contour plot with the combined effect of income and inhabitant sizes on the predicted value of soil sealing ego-hoods with a 1000 meters buffer.[44] Comparable to the other figures before, it consists of two subplots showing the results for the German and the migrant group. The lines in each subplot depict borders between categorized values of soil sealing, which are displayed in the legend. The darker a line is, the higher would be the value on a possible Z-axis of a three-dimensional

43 A correlation analysis with the German Census 2011 yields a correlation coefficient of .403 ($p \leq .001$) between the number of inhabitants and the number of immigrants in 1 km^2 census cells.

44 For this analysis, only the results of the models with soil sealing in buffers of 1000 meters are shown because this model revealed the most specific pattern before. Choosing different geographic scales still shows similar results, but for lucidity reasons, these figures are not shown.

plane. Thus, choosing an arbitrary X (income) and Y (inhabitant size) value also yields the corresponding Z value (soil sealing). Ultimately, the plot was created using a similar regression model as before, only with the additional three-way interactions between income, inhabitant size and migrant group status.



Note: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017); based on unimputed data; estimates are controlled for age, gender, education, homeownership, household size, number of inhabitants in municipality; N = 2,528

Figure 8.7: Predicted Values for Soil Sealing as a Function of Income and Municipality Inhabitant Sizes for German and Migrant People in Ego-Hoods with a 1000 Meters Buffer

What does the contour plot tell about the role of inhabitant sizes and soil sealing hazard inequalities among German and migrant people?

First, as inhabitant sizes are increasing, soil sealing are generally increasing as well. This positive correlation is as expected because land use indicators relate to population density. The higher the population density, the more likely more housings are built in a specific area—thus, the use of land increases, and so does the soil sealing.

Second, this relationship exists among German and migrant people, yet it is more pronounced for the German group. While also for the migrant group soil sealing increases with the inhabitant size of the municipality, it affects members of migrant groups with lower income even in smaller municipalities stronger than their German counterparts.

Third, the trajectories of soil sealing along the axis of income differ severely be-

tween German and migrant people. While the trajectories for the German group reflect the moderate decreasing slope that was shown before in Figure 8.5, they have a slight tendency to blur in high income groups combined with high inhabitant numbers. Nonetheless, the slope remains steeper for the migrant group.

Generally, the results corroborate the findings from before but introduce another perspective. Differences in soil sealing exposure among German and migrant people persist, and income cannot solely explain these differences. Instead, people with a migration background have to invest more income to live in housings with lower levels of soil sealing exposure. Also, introducing inhabitant sizes does not explain all aspects of this relationship—it emphasizes some of the patterns.

8.5 Discussion

Land use hazards by soil sealing affect people with different income and migration backgrounds on different levels. Income can help to reduce the general soil sealing exposure, which indicates that income helps people to afford better housings. Moreover, while soil sealing affects migrant people with low-income, at least some of them can reduce exposure with their income as well. However, this general relationship does not hold for these groups on all geographic levels. The relationship between income, migrant group status, and soil sealing exposure remains complicated.

Because of this complicated relationship, a final answer relating to the general mechanism in place is hard to give. Neither the proposed socio-economic theories nor the social processes theories provide clear explanations. Indeed, income eliminates some of the ethnic inequalities in the sample. Migrant people can significantly reduce soil sealing exposure with their income, which is indicative of the Ethnic Income Inequalities Thesis. At the same time, soil sealing hazards on some geographic levels remain intact—migrant people cannot approximate German's low levels of soil sealing with their income, which speaks for enduring ethnic discrimination.

The study of environmental and land use hazards, in particular, is as complicated as any other study of segregation or migration. Land use is a product of decisions made by people, institutions, and policymakers in the past, present, and the future. These actors control who lives where, and how places develop in the future—either on their own, on purpose, or as a result of social arrangements. Because of this multi-actor and longitudinal nature of the process, studying land use hazards as in this empirical application can only provide a snapshot of a specific situation at a specific time. Motives behind any action, the influence of past developments and social processes cannot directly be observed from the data points at hand.

Even so, what these data points suggest is that the general theories on environmental inequalities only partly apply to the German societal context. Predictions of the Socio-Economic or Ethnic Income Inequality Thesis may apply to some degree, but not entirely. It remains unclear whether this is a result of spatial assimilation or discrimination, both mechanisms would be in line with findings of others (Auspurg et al., 2017; Lersch, 2013). Generally, future work needs to concentrate on the integration of competing theories, and apply them with regards to the specific social context of migrant people.

8.5.1 Taking Migration History and Culture Into Account

Different migrant groups have different migration histories. This application's measure of migration background is agnostic about this fact as it considers the citizenship of the fathers of the survey respondents. Actual citizenship might get confused with cultural backgrounds by the respondents. There is a chance that for some people this operationalization does not apply and that the overall number of observations within the migration group is either over- or underestimated. Also, the effects of the whole group may be confounded with migration history.

The history of people with Russian parents in Germany, for example, is more complicated. Some of them, who are called Spätaussiedler, are people whose ancestors were rooted in the former Eastern regions of the German Empire and who lived in what was Soviet Russia after World War II. Many of the Spätaussiedler settled back to Germany in the late 1960ies and 1970ies and received the German citizenship, but their actual way of life differed in contrast to the general German population. Other German people at that time, in fact, perceived them as being foreign (Dirksmeier, 2014, 841), however, with regards to their actual citizenship they were not. Although other authors have challenged this argument by arguing that Spätaussiedler are nowadays perceived as a well-integrated group (Schaeffer, Höhne, & Teney, 2016), their migration history is still different from other groups. Confusing these people with people originating from other regions of Russia in a survey sample may produce confounding results.

On a different note, people with different migration backgrounds behave differently, thus, studying their moving behavior and motives becomes essential. Past research showed, for example, that Turkish people use green areas such as parks in a different way than the native population does (Kabisch & Haase, 2014, 131). The former often use them to arrange social events, such as celebrations and barbecues, and less for separate recreational activities. Thus, being less exposed to environmental

hazards such as soil sealing may not be part of their motive to move to other neighborhoods (Lersch, 2013).

8.5.2 Considering the Longitudinal Dimension

Studies that took into account the longitudinal dimension of exposure to environmental hazards partly contradict the findings of this application. After controlling for other person-level sociodemographic factors such as education, Best and Rüttenauer (2018) only found a weak link between income and, in their case, subjective air pollution. "[W]hen moving, households can use their income to reduce their exposure to pollution" (Best & Rüttenauer, 2018, 57), but this effect was not substantially significant. It even lost its statistical significance after including the citizenship of the respondents in the analysis.

Similarly, the effects in this study are rather small. Income reduces the level of land use hazards in the form of soil sealing just by a few percentage points. On the other hand, as soil sealing is just one of many stressors intruding people's lives, despite the sole small influence it can still be considered as being substantially significant. Newer research suggested that cumulative and combined exposure to environmental hazards increases the risk of their deleterious effects on people (Oiamo et al., 2015; Pedersen, 2015). Likewise, land use hazards are one of many correlated risk factors that affect people in their daily lives. Nevertheless, future research should also concentrate on the longitudinal dimension of the objectively measured indicators of land use hazards.

8.5.3 Limitations

The present application has some limitations worth noting. First, the sample size of the migrant group is comparably small. Statistical inferences lack statistical power and result in large proportions of uncertainty of the estimates. For this reason, the used methods remain basic, and more complicated models to compare both groups, such as SEM, are not possible to use. Second, the small sample size can also be an artifact of sample selection. The GESIS Panel is a survey conducted among the general population of Germany in the German language. This design does not consider all members of the migrant group because not all of them can speak the German language. Including these people in future research should lead to more pronounced effects. Third, the indicators of soil sealing are an average measure of air and watertight coverage of soils within an ego-hood and do not account for how neighborhoods relate to each other. However, previous research indicated that the constellation of neigh-

borhoods within a network of neighborhoods matters (Klinger et al., 2017; Legewie & Schaeffer, 2016), primarily relating to environmental hazards (Rüttenauer, 2018). Studies that incorporate such effects, combined with the statistical power of increased sample sizes and longitudinal data, may reveal strong patterns of land use hazard exposure. The analysis of inhabitant sizes on the municipality level already showed that geography matters.

8.5.4 Conclusion

Overall, the results of this application suggest that inequalities in land use hazard exposure exist in Germany. Land use hazards, such as soil sealing, affect people with low income and members of the migrant group more frequently than people with high income and members of the German population. Given the continually growing land use in Germany, these environmental inequalities may even increase in the next years. When cities aim to enhance the number of recreational areas to compensate for the deleterious effects of land use hazards (Kabisch & Haase, 2014), they should also take into account the inequality dimension of such efforts (Wolch et al., 2014). Thus, to find out which mechanisms are in place when people are moving should also be of high interest to policymakers.

9 Conclusion

This book deals with the use of georeferenced data for social science survey research which builds upon survey data enriched with direct spatial identifiers, such as geo-coordinates. It reviews the prerequisites and challenges of applying these data for different social science research questions, highlighting the different branches of an interdisciplinary effort. At the center of this presentation is the method of spatial linking: the combination of georeferenced survey data with information from auxiliary geospatial data sources. A collection of spatial linking methods is used in this book's empirical applications that underline these methods' flexibility in different social science sub-disciplines, such as family and health, political attitudes, and environmental inequalities research. In sum, locating survey respondents in space with georeferenced survey data opens up new avenues for research and innovation in the social sciences.

Researchers can break new ground and introduce innovation because they take a closer look at the context of social behavior and attitudes on a smaller scale than before. Georeferenced survey data based on respondents' addresses, in theory, are independent of common spatial classifications, such as municipalities or other administrative districts. Accordingly, researchers can add information from auxiliary geospatial data sources which do not need to rely on such classifications—geo-coordinates offer far more flexible methods for data linking. Two geospatial data sources of this book, for example, are road traffic noise data that are measured at the respondents' dwelling facade or rates of soil sealing on a level of 100 meters $\times$ 100 meters. Both sources are spatially linked one-to-one by their location but also by more sophisticated techniques, like for example geodesic distances and spatial buffers. In some ways, social science researchers are closer to the respondents than before, which makes asking new research questions or reassessing old ones possible in the first place.

The first empirical application in Chapter 6 effectively expands on existing work in family and health research. Other authors have shown that married people in Western societies are still better off with regards to their health when it comes to external stress intruding their lives. It is the quality of social ties, social influences on stress processing, and the effectiveness of dyadic coping which explain why the effects of environmental stressors, such as road traffic noise, are potentially buffered in marriages. After spatially linking data from the georeferenced GGSS with road traffic noise data, the analysis, using an SEM approach, partly corroborates this hypothesis. In contrast to married people, unmarried people report more mental health strains when they are exposed to road traffic noise in their home. While this relationship is

more complicated in detail—for example, married people report more noise annoyance after road traffic noise exposure—, this empirical application provides external validity to findings from previous studies: married people show less deleterious responses to stressors. Without the use of georeferenced survey data spatially linked to road traffic noise data, establishing external validity would not have been possible.

The second empirical application in Chapter 7 contributes to the vast body of research on ethnic diversity and prejudices. For a long time, scholars have discussed whether, for example, immigrant people in the neighborhood decrease or increase the likelihood of developing xenophobic attitudes among native-born inhabitants, and what the exact mechanisms may be. Different theories, such as the Contact Theory and the Intergroup Threat or Ethnic Competition Theory, provide plausible explanations why some people develop xenophobic attitudes and others do not. Confusingly, while they provide competing explanations, there has been evidence for all of them in various publications. The Halo Hypothesis, tested in this empirical application, provides a synthesis to solve this puzzle: it states that people who live in ethnically homogeneous neighborhoods which border ethnically diverse neighborhoods are more probable to develop xenophobic attitudes. Using the data from the georeferenced GGSS spatially linked to detailed information on immigrant rates in direct and surrounding neighborhoods, and applying an extensive set of operationalizations and estimations, the analysis, however, reveals no evidence for a Halo effect in Germany. While these results are unexpected given the evidence in other countries such as Switzerland or Sweden, they demonstrate how sensitive theories are to differences in societal contexts, in this case, to differences in ethnic segregation. Without the use of georeferenced survey data, such detailed insights into the spatial integration of social science theories would not have been achieved.

The final empirical application in Chapter 8 complements the research on environmental inequalities which is surprisingly understudied in Germany. Generally, environmental inequalities research is involved with inequalities between distinct social groups and their exposure to environmental hazards, such as air pollution or road traffic noise. However, only in the last few years, more and more researchers have tried to transfer work of scholars who studied the US societal context to the German societal context, an effort which this empirical application also aims to accomplish. It is asked how the income of people and their migrant status relate to environmental hazards originating in soil sealing. After spatially linking georeferenced survey data from the GESIS Panel to the soil sealing data, a series of linear prediction models yields clear patterns of ethnic inequalities which, besides some income differences, may also be caused by ethnic discrimination. People who have a migrant background

are at higher risk to be exposed to soil sealing—the risk only decreases in high-income groups. Moreover, a robustness checks reveals that these findings are not confounded by higher rates of migrant people in larger communities in which soil sealing is more prevalent than in smaller communities. Once more, this empirical application exemplifies that without the use of georeferenced survey data, the exploration of these patterns on such a fine-grained geographic level could not have been accomplished.

Overall, these empirical applications in the fields of health and family, political attitudes as well as environmental inequalities display the variety of social science subdisciplines that can profit from using georeferenced data. Thus far, they also present an emerging field of research for social scientists, requiring new analytic skills from diverse and foreign disciplines, such as ecology and engineering. Navigating the organizational and technical requirements for the analysis of georeferenced survey data enables researchers to answer new and innovative research questions. This book's research highlights some of these questions in the studies of road traffic noise impacts on health and the moderating role of marriage, inter-neighborhood effects between immigrant rates and xenophobia, and migrants' land use hazard exposure as a function of income.

9.1 Combining Data from Different Domains

As noted throughout this book, using georeferenced survey data and spatial linking methods is an interdisciplinary discovery (Dietz, 2002). What makes it such a challenge is that the data from different scientific domains are combined and processed with the theoretical background from one discipline and the methodological considerations of another. For example, Chapter 8 draws on classic social science theories about social inequalities but applies data from the sciences of landscape planning. In contrast to survey measures, the land use indicator of soil sealing had to be justified as an appropriate measure of an environmental hazard because it is foreign to standard social science research. Accordingly, researchers aiming to conduct such research require information about the methods and data of the social sciences *and* of the disciplines from which the geospatial information originates.

One of the most apparent issues of combining both data sources is that they were not collected for the same purpose (Parsons et al., 2011). Survey projects often do not take samples in order to draw conclusions from the sample about the people within a small region but for the population of a whole country, or at least larger regions. In this setting, small geographic units, such as municipalities, comprise sample points which have a specific probability to get drawn. They contribute to a preferably representative

picture of the population (Zhang, 2012, 42) but not for their sample point. Some of the geospatial data, in contrast, aim to provide an extensive picture of particular small geographic areas, such as 100 meters × 100 meters grid cells. As in the case of the road traffic noise data, they offer a fine-grained level of detail for a specific point on a map but are censored below a specific level—50 dB(A) in the case of the road traffic noise data. In either case, survey data and geospatial data stem from distinct modes of collection and differ in their data generation process.

Because of these different data generation processes, the choice of a specific spatial linking procedure cannot be derived directly from a short glimpse at the data. Researchers have to develop hypotheses about how measures of geospatial data affect people in the same geographic space, and they have to operationalize concepts into the proper choice of an GIS method (see Chapter 5.2). What adds to this issue is that social science theories are seldom specific with regards to such hypotheses, which makes research with georeferenced survey data also to an exploratory maneuver. For instance, the term neighborhood is often used as a rather broad term (Sampson et al., 2002) which researchers apply to describe small areas of a few hundred square meters (Perchoux et al., 2016) or large areas, such as municipalities (Hanson & Hawley, 2011). To make a methodological contribution to this literature, recent studies using small-scale geospatial data vary the size of neighborhoods in order to detect the contextual effects that may be most relevant for people (see, e.g., Klinger et al., 2017; Sluiter et al., 2015; Tolsma & van der Meer, 2017).

In sum, researchers who work at this interface of different scientific disciplines do not only need the domain knowledge of each discipline, but they also need to translate it from one discipline to another. Some research questions involve econometric approaches such as the deployment of instrumental variables based on measures gathered from an GIS method (Chapter 6.2.1). Other research questions require to analyze relationships between different geographic units and to transfer the results to the individual survey respondent level (Chapter 7.2.2). Alternatively, as some geospatial data are available on a particular small geographic scale, some research questions make it necessary to vary the size of neighborhoods to identify, as noted above, the relevant contextual effects (Chapter 8.2.2). All these efforts require a thorough knowledge of both data sources.

9.2 Geospatial Data as Contextual Data

The social sciences are largely about the context of social behavior and attitudes, emphasizing how external factors affect people's lives. Generations of researchers al-

ready have harnessed contextual data which differ from individual data by primarily containing information not produced by individual people. What researchers have aimed to explain with such data are behavior and attitudes of people who are exposed to different economic settings (Dixon, Fullerton, & Robertson, 2013; Otjes & Katsanidou, 2017), policies (Andrews & Seguin, 2015; Brandt & Deindl, 2013; You et al., 2011), infrastructures (Finke & Adamczyk, 2008; Kjellstrom et al., 2007), or cultural environments (Falster et al., 2016; Raijman et al., 2008). The data gained from spatial linking projects with georeferenced survey data are not different in this regard as spatial information adds context to existing individual data.

Georeferenced survey data and spatial linking, however, differ in the context's level of detail. As mentioned, social scientists are closer to the respondent than before, which opens up new opportunities to decompose individual and contextual effects. While it remains questionable whether effects are contextual at all or just a composition of individual characteristics (Ross & Mirowsky, 2008), which in some cases can already be answered with data gathered on the regional level (Hank & Huinink, 2015), a perspective on small-scale neighborhoods nonetheless provides new vehicles to research. Georeferenced survey data allow to add information which varies on a small geographic scale, such as environmental data, and they offer to assess research questions that target small neighborhood dynamics, such as intra-neighborhood migration. Conceptually, georeferenced survey data provide similar research opportunities as common contextual data, but they also offer new information and an additional toolbox of methods to accomplish this effort.

9.3 Gain in Knowledge by Using Georeferenced Survey Data

Within this emerging field researchers are confronted with questions of how these additional data build upon existing knowledge within the social sciences. For example, in Chapter 6 of this book, indications are found that married people experience less deleterious effects from exposure to road traffic noise stressors than unmarried people. These effects are less pronounced than other stress measures in the literature, e.g., work stress (Sandberg et al., 2013), but they give external validity to ongoing debates in the social sciences. A finding from Chapter 7 yields null-effects of neighborhood immigrant rates on xenophobia. These results are unexpected, but they illustrate that also small-scale contextual effects are not independent of societal contexts, such as different manifestations of ethnic segregation. All of these effects are small, but they support, complement, and extend existing research.

These findings thus impressively illustrate what the gain in knowledge from the

use of georeferenced survey data is. A promise of applying them to contemporary research questions may be that they reveal better insights into the context of social behavior and attitudes (Stimson, 2014), but this does not mean that the results are automatically stronger, more pronounced, or even more statistically significant (Dietz, 2002). Indeed, researchers can expect to be better prepared for logical traps, such as the ecological fallacy (Chapter 3.2.3); however, classification and, more importantly, explanations of results are still needed. Neighborhood effects involve complicated relationships between people's migration, their selection of specific areas or the siting of infrastructures, such that any found effect always depicts a stationary result that needs unraveling of these factors.

9.4 Outlook

This book's presentation of georeferenced data in social science survey research hardly comprises an exhaustive treatise about the "spatially integrated social science" (Goodchild et al., 2000). There are plenty of ongoing developments in the field which are not covered here, e.g., the use of sensor data to retrieve geo-coordinates from the *Global Positioning System* (GPS) automatically (Bluemke et al., 2017; Kamilaris & Ostermann, 2018). Some of these developments are helpful as they ease to navigate challenges, for example, by promising that researchers no longer have to rely on third-party service providers for geocoding in the case of sensor data. Using modern technologies as a tool in surveys helps to make geospatial information to an integral part of research.

Moreover, methods and concepts can only be widely used and standardized if data are accessible. Many of the neighborhood measures used by researchers are exploratory, stemming from data that are at hand and therefore are not theoretically justified. Alternatively, neighborhood operationalizations are a result of try and error to find relevant context effects, e.g., by varying buffer sizes. This book's empirical applications are not an exception, as in Chapter 7 and Chapter 8 varying sizes of neighborhoods are used. The field, however, can only move on when a broad range of researchers gets access to the data to validate measures from exploratory findings with confirmatory analyses. What is needed is widespread access to georeferenced data so that the scientific community can jointly develop measures of neighborhood effects, comparable to standardized items in survey research.

As indicated, also the data landscape is changing. This change does not only occur in the relation between public data and commercial data but also with regards to newly available data types, originating from a diverse set of sources. Data from the inter-

net, such as social media data from Twitter (Wang, Phillips, Small, & Sampson, 2018), Facebook (Müller & Schwarz, 2017) and others (Hristova, Aiello, & Quercia, 2018), also open up new avenues for conducting research. Not all of these data are georeferenced data, but sometimes they provide new research opportunities for other domains of the social sciences, such as network analysis (Wimmer & Lewis, 2010). Some of these sources at least have a chance to provide indirect location information that can be geocoded (Stefanidis, Crooks, & Radzikowski, 2013). Generally, as the data landscape is changing rapidly, also social science survey researchers who intend to link their data with other sources do well to make themselves familiar with current and future developments.

To conclude, if social science research aims to explain social phenomena, only theory can be the answer to disentangle the snares and challenges of the data landscape. Data are the vehicles to test hypotheses (Stimson, 2014), but data never stand for themselves. Also, working with georeferenced survey data only makes sense if researchers know the meaning of geo-coordinates—they enable researchers to operationalize survey respondents' direct living environment. Consequently, georeferenced survey research profits from applications that assess causal questions, e.g., by using longitudinal data (Crowder et al., 2011; Fecht et al., 2016), and researchers who reconsider measurements of neighborhoods, such as the works on effective neighborhoods (Spielman & Yoo, 2009), ego-hoods (Perchoux et al., 2016) and t-communities (Foster & Hipp, 2011).[45] Future studies should focus on the spatial integration of theories, so that they incorporate specific hypotheses on space and social phenomena. Nonetheless, such projects can only be managed by concatenating theory and data tightly.

45 t-communities are discussed as providing a flexible method within GIS to define "geospatial boundaries that respect the logic of social interaction" (Foster & Hipp, 2011, 27). In contrast to buffer methods, for example, they adjust the neighborhood measure of natural boundaries (see Chapter 5.2.2).

References

Abadie, A., Athey, S., Imbens, G., & Wooldridge, J. (2017, November). *When Should You Adjust Standard Errors for Clustering?* (Tech. Rep. No. w24003). Cambridge, Massachusetts: National Bureau of Economic Research. doi: 10.3386/w24003

Ainsworth, J. W. (2002, September). Why Does It Take a Village? The Mediation of Neighborhood Effects on Educational Achievement. *Social Forces, 81*(1), 117–152. doi: 10.1353/sof.2002.0038

Allport, G. W. (1954). *The Nature of Prejudice.* Cambridge, Massachusetts: Addison-Wesley Publishing Company.

Andrews, K. T., & Seguin, C. (2015, September). Group Threat and Policy Change: The Spatial Dynamics of Prohibition Politics, 1890–1919. *American Journal of Sociology, 121*(2), 475–510. doi: 10.1086/682134

Aneshensel, C. S. (2015, June). Sociological Inquiry into Mental Health: The Legacy of Leonard I. Pearlin. *Journal of Health and Social Behavior, 56*(2), 166–178. doi: 10.1177/0022146515583992

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996, June). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association, 91*(434), 444–455. doi: 10.1080/01621459.1996.10476902

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Auspurg, K., Hinz, T., & Schmid, L. (2017, March). Contexts and Conditions of Ethnic Discrimination: Evidence from a Field Experiment in a German Housing Market. *Journal of Housing Economics, 35*, 26–36. doi: 10.1016/j.jhe.2017.01.003

Babisch, W. (2014). Updated Exposure-Response Relationship Between Road Traffic Noise and Coronary Heart Diseases: A Meta-Analysis. *Noise and Health, 16*(68), 1–9. doi: 10.4103/1463-1741.127847

Babisch, W., Wolf, K., Petz, M., Heinrich, J., Cyrys, J., & Peters, A. (2014, March). Associations between Traffic Noise, Particulate Air Pollution, Hypertension, and Isolated Systolic Hypertension in Adults: The KORA Study. *Environmental Health Perspectives, 122*(5), 492–498. doi: 10.1289/ehp.1306981

Banton, M. (1983). *Racial and Ethnic Competition*. Cambridge & New York: Cambridge University Press.

Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., & Stansfeld, S. (2014, April). Auditory and Non-Auditory Effects of Noise on Health. *The Lancet, 383*, 1325–1332. doi: 10.1016/S0140-6736(13)61613-X

Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the Left-Right Scale

a Valid Measure of Ideology? *Political Behavior*, *39*(3), 553–583. doi: 10.1007/s11109-016-9368-2

Baur, N., Hering, L., Raschke, A. L., & Thierbach, C. (2014). Theory and Methods in Spatial Analysis. Towards Integrating Qualitative, Quantitative and Cartographic Approaches in the Social Sciences and Humanities. *Historical Social Research*, *39*(2), 7–50. doi: 10.12759/hsr.39.2014.2.7-50

Bensmann, F., Zapilko, B., & Mayr, P. (2017, March). Interlinking Large-scale Library Data with Authority Records. *Frontiers in Digital Humanities*, *4*, 1–13. doi: 10.3389/fdigh.2017.00005

Berkman, L. F., Glass, T., Brissette, I., & Seeman, T. E. (2000, September). From Social Integration to Health: Durkheim in the New Millennium. *Social Science & Medicine*, *51*(6), 843–857. doi: 10.1016/S0277-9536(00)00065-4

Best, H., & Rüttenauer, T. (2018, February). How Selective Migration Shapes Environmental Inequality in Germany: Evidence from Micro-level Panel Data. *European Sociological Review*, *34*(1), 52–63. doi: 10.1093/esr/jcx082

Bivand, R., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. New York: Springer.

Blalock, H. M. (1967). *Toward a Theory of Minority-Group Relations*. New York: Wiley.

Blatt, A. J. (2012, January). Ethics and Privacy Issues in the Use of GIS. *Journal of Map & Geography Libraries*, *8*(1), 80–84. doi: 10.1080/15420353.2011.627109

Bluemke, M., Resch, B., Lechner, C., Westerholt, R., & Kolb, J.-P. (2017). Integrating Geographic Information into Survey Research: Current Applications, Challenges and Future Avenues. *Survey Research Methods*, *11*(3), 307–327. doi: 10.18148/srm/2017.v11i3.6733

Blumer, H. (1958, April). Race Prejudice as a Sense of Group Position. *The Pacific Sociological Review*, *1*(1), 3–7. doi: 10.2307/1388607

Bocquier, A., Cortaredona, S., Boutin, C., David, A., Bigot, A., Chaix, B., … Verger, P. (2013, August). Small-Area Analysis of Social Inequalities in Residential Exposure to Road Traffic Noise in Marseilles, France. *The European Journal of Public Health*, *23*(4), 540–546. doi: 10.1093/eurpub/cks059

Bocquier, A., Cortaredona, S., Boutin, C., David, A., Bigot, A., Sciortino, V., … Verger, P. (2014, April). Is Exposure to Night-Time Traffic Noise a Risk Factor for Purchase of Anxiolytic-Hypnotic Medication? A Cohort Study. *The European Journal of Public Health*, *24*(2), 298–303. doi: 10.1093/eurpub/ckt117

Bodenmann, G. (1997). Dyadic Coping: A Systemic-Transactional View of Stress and Coping Among Couples: Theory and Empirical Findings. *European Review of Applied Psychology*, *47*(2), 137–141.

Boes, S., Nüesch, S., & Stillman, S. (2013, September). Aircraft Noise, Health, and Residential Sorting: Evidence from Two Quasi-Experiments: Aircraft Noise and Health. *Health Economics*, *22*(9), 1037–1051. doi: 10.1002/hec.2948

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel. *Social Science Computer Review*, *36*(1), 103–115. doi: 10.1177/0894439317697949

Bowyer, B. (2008). Local Context and Extreme Right Support in England: The British National Party in the 2002 and 2003 Local Elections. *Electoral Studies*, *27*(4), 611–620. doi: 10.1016/j.electstud.2008.05.001

Brandt, M., & Deindl, C. (2013, February). Intergenerational Transfers to Adult Children in Europe: Do Social Policies Matter? *Journal of Marriage and Family*, *75*(1), 235–251. doi: 10.1111/j.1741-3737.2012.01028.x

Braubach, M., & Fairburn, J. (2010, February). Social Inequities in Environmental Risks Associated with Housing and Residential Location-a Review of Evidence. *European Journal of Public Health*, *20*(1), 36–42. doi: 10.1093/eurpub/ckp221

Brink, M. (2011, May). Parameters of Well-Being and Subjective Health and Their Relationship with Residential Traffic Noise Exposure - a Representative Evaluation in Switzerland. *Environment International*, *37*(4), 723–733. doi: 10.1016/j.envint.2011.02.011

Brüderl, J., Hank, K., Huinink, J., Nauck, B., Neyer, F. J., Walper, S., … Wilhelm, B. (2015). *The German Family Panel (pairfam).* GESIS Data Archive. doi: 10.4232/pairfam.5678.6.0.0

Carr, D., & Springer, K. W. (2010, June). Advances in Families and Health Research in the 21st Century. *Journal of Marriage and Family*, *72*(3), 743–761. doi: 10.1111/j.1741-3737.2010.00728.x

Chhetri, P., & Stimson, R. J. (2014). Merging Survey and Spatial Data Using GIS-Enabled Analysis and Modelling. In R. J. Stimson (Ed.), *Handbook of Research Methods and Applications in Spatially Integrated Social Science* (pp. 511–534). Cheltenham, Northampton: Edward Elgar Publishing Inc.

Christakis, N. A., & Fowler, J. H. (2013, February). Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior. *Statistics in Medicine*, *32*(4), 556–577. doi: 10.1002/sim.5408

City of Cologne. (2014). *City Districts Shapefiles.* Retrieved September 28, 2018, from https://offenedaten-koeln.de/sites/default/files/Stadtteil.zip.

Coan, J. A., Schaefer, H. S., & Davidson, R. J. (2006, December). Lending a Hand: Social Regulation of the Neural Response to Threat. *Psychological Science*, *17*(12),

1032–1039. doi: 10.1111/j.1467-9280.2006.01832.x

Cohen, J. (1994). The Earth is Round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Crowder, K., & Downey, L. (2010). Inter-Neighborhood Migration, Race, and Environmental Hazards: Modeling Micro-Level Processes of Environmental Inequality. *American Journal of Sociology*, *115*(4), 1110–1149.

Crowder, K., Hall, M., & Tolnay, S. E. (2011). Neighborhood Immigration and Native Out-Migration. *American Sociological Review*, *76*(1), 25–47. doi: 10.1177/0003122410396197

Crowder, K., & South, S. J. (2011, January). Spatial and Temporal Dimensions of Neighborhood Effects on High School Graduation. *Social Science Research*, *40*(1), 87–106. doi: 10.1016/j.ssresearch.2010.04.013

Curry, A., Latkin, C., & Davey-Rothwell, M. (2008, July). Pathways to Depression: The Impact of Neighborhood Violent Crime on Inner-City Residents in Baltimore, Maryland, USA. *Social Science & Medicine*, *67*(1), 23–30. doi: 10.1016/j.socscimed.2008.03.007

Cutrona, C. E., Wallace, G., & Wesner, K. A. (2006, August). Neighborhood Characteristics and Depression: An Examination of Stress Processes. *Current Directions in Psychological Science*, *15*(4), 188–192. doi: 10.1111/j.1467-8721.2006.00433.x

Dean, N., Dong, G., Piekut, A., & Pryce, G. (2018, April). Frontiers in Residential Segregation: Understanding Neighbourhood Boundaries and Their Impacts. *Tijdschrift voor Economische en Sociale Geografie*, 1–18. doi: 10.1111/tesg.12316

De Smith, M. J., Goodchild, M. F., & Longley, P. (2018). *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*.

Diekmann, A., & Meyer, R. (2010, September). Demokratischer Smog? Eine empirische Untersuchung zum Zusammenhang zwischen Sozialschicht und Umweltbelastungen. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *62*(3), 437–457. doi: 10.1007/s11577-010-0108-z

Dietz, R. D. (2002, December). The Estimation of Neighborhood Effects in the Social Sciences: An Interdisciplinary Approach. *Social Science Research*, *31*(4), 539–575. doi: 10.1016/S0049-089X(02)00005-4

Dill, V., & Jirjahn, U. (2014, December). Ethnic Residential Segregation and Immigrants' Perceptions of Discrimination in West Germany. *Urban Studies*, *51*(16), 3330–3347. doi: 10.1177/0042098014522719

Dinesen, P. T., & Sønderskov, K. M. (2015, June). Ethnic Diversity and Social Trust: Evidence from the Micro-Context. *American Sociological Review*, *80*(3), 550–573. doi: 10.1177/0003122415577989

Dirksmeier, P. (2014, November). Are Urbanites More Permissive? Germany's Ur-

ban Geography of Prejudice. *Urban Affairs Review, 50*(6), 835–863. doi: 10.1177/1078087414520950

Dixon, J. C., Fullerton, A. S., & Robertson, D. L. (2013, October). Cross-National Differences in Workers' Perceived Job, Labour Market, and Employment Insecurity in Europe: Empirical Tests and Theoretical Extensions. *European Sociological Review, 29*(5), 1053–1067. doi: 10.1093/esr/jcs084

Downey, L., Crowder, K., & Kemp, R. J. (2016, September). Family Structure, Residential Mobility, and Environmental Inequality: Family Structure and Environmental Inequality. *Journal of Marriage and Family, 79*(2), 535–555. doi: 10.1111/jomf.12355

Drefahl, S. (2012, June). Do the Married Really Live Longer? The Role of Cohabitation and Socioeconomic Status. *Journal of Marriage and Family, 74*(3), 462–475. doi: 10.1111/j.1741-3737.2012.00968.x

El Emam, K. (2006). *Overview of Factors Affecting the Risk of Re-identification in Canada* (Tech. Rep.).

Elhorst, J. P. (2014). Linear Spatial Dependence Models for Cross-Section Data. In *Spatial Econometrics* (pp. 5–36). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-40340-8_2

Eriksson, C., Hilding, A., Pyko, A., Bluhm, G., Pershagen, G., & Östenson, C.-G. (2014, May). Long-Term Aircraft Noise Exposure and Body Mass Index, Waist Circumference, and Type 2 Diabetes: A Prospective Study. *Environmental Health Perspectives, 122*(7), 687–694. doi: 10.1289/ehp.1307115

ESRI. (1998). *ESRI Shapefile Technical Description* (Tech. Rep.).

ESRI. (2015). *ArcGIS Desktop: Release 10.3.* ESRI - Environmental Systems Research Institute. Redlands, California.

European Parliament, & European Council. (2002). *Directive 2002/49/EC of the European Parliament and of the European Council.*

European Parliament, & European Council. (2008). *Regulation (EC) No 763/2008 on Population and Housing Censuses.*

Falster, K., Banks, E., Lujic, S., Falster, M., Lynch, J., Zwi, K., … Jorm, L. (2016, December). Inequalities in Pediatric Avoidable Hospitalizations Between Aboriginal and Non-Aboriginal Children in Australia: A Population Data Linkage Study. *BMC Pediatrics, 16*(1), 1–12. doi: 10.1186/s12887-016-0706-7

Fecht, D., Hansell, A. L., Morley, D., Dajnak, D., Vienneau, D., Beevers, S., … Gulliver, J. (2016, March). Spatial and Temporal Associations of Road Traffic Noise and Air Pollution in London: Implications for Epidemiological Studies. *Environment International, 88*, 235–242. doi: 10.1016/j.envint.2015.12.001

Federal Agency for Cartography and Geodesy. (2018). *Administrative Areas 2018 Shapefiles.* Retrieved May 29, 2018, from https://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div?gdz_spr=eng&gdz_akt_zeile=5&gdz_anz_zeile=1&gdz_unt_zeile=14&gdz_user_id=0.

Federal Agency for the Environment Germany. (2015). *Data on the Environment 2015*.

Finke, R., & Adamczyk, A. (2008, September). Cross-National Moral Beliefs: The Influence of National Religious Context. *The Sociological Quarterly, 49*(4), 617–652. doi: 10.1111/j.1533-8525.2008.00130.x

Förster, A. (2018, June). Ethnic Heterogeneity and Electoral Turnout: Evidence from Linking Neighbourhood Data with Individual Voter Data. *Electoral Studies, 53*, 57–65. doi: 10.1016/j.electstud.2018.03.002

Foster, K. A., & Hipp, J. A. (2011, March). Defining Neighborhood Boundaries for Social Measurement: Advancing Social Work Research. *Social Work Research, 35*(1), 25–35. doi: 10.1093/swr/35.1.25

Gelman, A. (2006). Prior Distributions for Variance Parameters in Hierarchical Models. *Bayesian Analysis, 1*(3), 515–533.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

German Environmental Agency / EIONET Central Data Repository. (2016). *Road Traffic Noise 2012 Shapefiles.* Retrieved November 30, 2016, from https://github.com/stefmue/georefum/blob/master/data/cdr.road.lden.dat.rda.

GESIS - Leibniz Institute for the Social Sciences. (2015). *ALLBUS/GGSS 2014 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2014).* GESIS Data Archive. doi: 10.4232/1.12209

GESIS - Leibniz Institute for the Social Sciences. (2017). *GESIS Panel - Extended Edition.* GESIS Data Archive. doi: 10.4232/1.12742

GESIS - Leibniz Institute for the Social Sciences. (2018). *ALLBUS/GGSS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey) - Sensitive Regional Data.* GESIS Data Archive. doi: 10.4232/1.13010

Gibbons, S., & Overman, H. G. (2012, May). Mostly Pointless Spatial Econometrics? *Journal of Regional Science, 52*(2), 172–191. doi: 10.1111/j.1467-9787.2012.00760.x

Gidlöf-Gunnarsson, A., & Öhrström, E. (2007, November). Noise and Well-Being in Urban Residential Environments: The Potential Role of Perceived Availability to Nearby Green Areas. *Landscape and Urban Planning, 83*(2-3), 115–126. doi: 10.1016/j.landurbplan.2007.03.003

Gleditsch, K. S., & Weidmann, N. B. (2012, June). Richardson in the Information Age: Geographic Information Systems and Spatial Data in International Studies.

Annual Review of Political Science, *15*(1), 461–481. doi: 10.1146/annurev-polisci-031710-112604

Goebel, J. (2017). *SOEP 2015 -Informationen zu den SOEP-Geocodes in SOEP v32* (Tech. Rep.). DIW Berlin, German Institute for Economic Research.

Goebel, J., & Hoppe, L. (2015). *Ausmaß und Trends sozialräumlicher Segregation in Deutschland*.

Goebel, J., Spieß, C. K., Witte, N. R. J., & Gerstenberg, S. (2014). *Die Verknüpfung des SOEP mit MICROM-Indikatoren: Der MICROM-SOEP-Datensatz* (Tech. Rep. No. 233). DIW Berlin, German Institute for Economic Research.

Goodchild, M. F., Anselin, L., Appelbaum, R. P., & Harthorn, B. H. (2000, April). Toward Spatially Integrated Social Science. *International Regional Science Review*, *23*(2), 139–159. doi: 10.1177/016001760002300201

Graham, J. E., Christian, L. M., Kiecolt-Glaser, J. K., & Ader, R. (2007). Close Relationships and Immunity. In *Psychoneuroimmunology* (Vol. 2, pp. 781–798). Burlington, Massachusetts: Elsevier.

Guite, H., Clark, C., & Ackrill, G. (2006, December). The Impact of the Physical and Urban Environment on Mental Well-Being. *Public Health*, *120*(12), 1117–1126. doi: 10.1016/j.puhe.2006.10.005

Haines, V. A., Beggs, J. J., & Hurlbert, J. S. (2011, March). Neighborhood Disadvantage, Network Social Capital, and Depressive Symptoms. *Journal of Health and Social Behavior*, *52*(1), 58–73. doi: 10.1177/0022146510394951

Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016, April). Current State of Linked Data in Digital Libraries. *Journal of Information Science*, *42*(2), 117–127. doi: 10.1177/0165551515594729

Hank, K., & Huinink, J. (2015). Regional Contexts and Family Formation: Evidence from the German Family Panel. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *76*(Supplement), 41–58. doi: 10.1007/s11577-015-0322-9

Hank, K., Jürges, H., & Schaan, B. (2009). Die Erhebung biometrischer Daten im Survey of Health, Ageing and Retirement in Europe. *Methoden – Daten – Analysen*, *3*(1), 97–108. doi: 10.2139/ssrn.1480195

Hanson, A., & Hawley, Z. (2011, September). Do Landlords Discriminate in the Rental Housing Market? Evidence from an Internet Field Experiment in US Cities. *Journal of Urban Economics*, *70*(2-3), 99–114. doi: 10.1016/j.jue.2011.02.003

Hardoy, M. C., Carta, M. G., Marci, A. R., Carbone, F., Cadeddu, M., Kovess, V., ... Carpiniello, B. (2005, January). Exposure to Aircraft Noise and Risk of Psychiatric Disorders: The Elmas Survey: Aircraft Noise and Psychiatric Disorders. *Social Psychiatry and Psychiatric Epidemiology*, *40*(1), 24–26. doi: 10.1007/

s00127-005-0837-x

Helbling, M. (2011). Why Swiss-Germans Dislike Germans: Opposition to Culturally Similar and Highly Skilled Immigrants. *European Societies, 13*(1), 5–27. doi: 10.1080/14616696.2010.533784

Hello, E., Scheepers, P., & Sleegers, P. (2006). Why the More Educated Are Less Inclined to Keep Ethnic Distance: An Empirical Test of Four Explanations. *Ethnic and Racial Studies, 29*(5), 959–985. doi: 10.1080/01419870600814015

Henry, K. A., & Boscoe, F. P. (2008). Estimating the Accuracy of Geographical Imputation. *International Journal of Health Geographics, 7:3*, 1–10. doi: 10.1186/1476-072X-7-3

Heritier, H., Vienneau, D., Frei, P., Eze, I. C., Brink, M., Probst-Hensch, N., & Roeoesli, M. (2014, December). The Association between Road Traffic Noise Exposure, Annoyance and Health-Related Quality of Life (HRQOL). *International Journal of Environmental Research and Public Health, 11*(12), 12652–12667. doi: 10.3390/ijerph111212652

Hijmans, R. J. (2017). *Introduction to the 'raster' Package (version 2.6-7)* (Tech. Rep.).

Hillmert, S., Hartung, A., & Weßling, K. (2017, October). Dealing with Space and Place in Standard Survey Data. *Survey Research Methods. Special Issue: Uses of Geographic Information Systems Tools in Survey Data Collection and Analysis, 11*(3), 267–287. doi: 10.18148/srm/2017.v11i3.6729

Hoffmann, B., Robra, B. P., & Swart, E. (2003, June). Social Inequality and Noise Pollution by Traffic in the Living Environment - an Analysis by the German Federal Health Survey (bundesgesundheitssurvey). *Gesundheitswesen, 65*(6), 393–401. doi: 10.1055/s-2003-40308

Hopkins, D. J. (2010, February). Politicized Places: Explaining Where and When Immigrants Provoke Local Opposition. *American Political Science Review, 104*(1), 40–60. doi: 10.1017/S0003055409990360

Hristova, D., Aiello, L. M., & Quercia, D. (2018, April). The New Urban Success: How Culture Pays. *Frontiers in Physics, 6*, 1–13. doi: 10.3389/fphy.2018.00027

Jakovljevic, B., Paunovic, K., & Belojevic, G. (2009, April). Road-Traffic Noise and Factors Influencing Noise Annoyance in an Urban Population. *Environment International, 35*(3), 552–556. doi: 10.1016/j.envint.2008.10.001

Janmaat, J. G. (2014). Do Ethnically Mixed Classrooms Promote Inclusive Attitudes Towards Immigrants Everywhere? A Study Among Native Adolescents in 14 Countries. *European Sociological Review, 30*(6), 810–822. doi: 10.1093/esr/jcu075

Johnson, M. D., Horne, R. M., & Galovan, A. M. (2016). The Developmental Course of Supportive Dyadic Coping in Couples. *Developmental Psychology, 52*(12), 2031–

2043. doi: 10.1037/dev0000216

Jurcik, T., Ahmed, R., Yakobov, E., Solopieieva-Jurcikova, I., & Ryder, A. G. (2013, August). Understanding the Role of the Ethnic Density Effect: Issues of Acculturation, Discrimination and Social Support: Ethnic Density and Acculturation. *Journal of Community Psychology, 41*(6), 662–678. doi: 10.1002/jcop.21563

Kabisch, N., & Haase, D. (2014, February). Green Justice or Just Green? Provision of Urban Green Spaces in Berlin, Germany. *Landscape and Urban Planning, 122*, 129–139. doi: 10.1016/j.landurbplan.2013.11.016

Kamilaris, A., & Ostermann, F. (2018, July). Geospatial Analysis and the Internet of Things. *ISPRS International Journal of Geo-Information, 7*(7), 1–22. doi: 10.3390/ijgi7070269

Kilpi, F., Konttinen, H., Silventoinen, K., & Martikainen, P. (2015, May). Living Arrangements as Determinants of Myocardial Infarction Incidence and Survival: A Prospective Register Study of Over 300,000 Finnish Men and Women. *Social Science & Medicine, 133*, 93–100. doi: 10.1016/j.socscimed.2015.03.054

Kim, J., & Ross, C. E. (2009). Neighborhood-Specific and General Social Support: Which Buffers the Effect of Neighborhood Disorder on Depression? *Journal of Community Psychology, 37*(6), 725–736. doi: 10.1002/jcop.20327

Kjellstrom, T., Friel, S., Dixon, J., Corvalan, C., Rehfuess, E., Campbell-Lendrum, D., ... Bartram, J. (2007, May). Urban Environmental Health Hazards and Health Equity. *Journal of Urban Health, 84*(S1), 86–97. doi: 10.1007/s11524-007-9171-9

Klinger, J. (2018). *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften - ALLBUS Sensitive Regionaldaten* (Tech. Rep. No. 2018|15). GESIS Data Archive. doi: 10.4232/1.13010

Klinger, J., Müller*, S., & Schaeffer, M. (2017). Der Halo-Effekt in einheimisch-homogenen Nachbarschaften: Steigert die ethnische Diversität angrenzender Nachbarschaften die Xenophobie? *Zeitschrift für Soziologie, 46*(6), 402–419. doi: 10.1515/zfsoz-2017-1022

Kohlhuber, M., Mielck, A., Weiland, S. K., & Bolte, G. (2006, June). Social Inequality in Perceived Environmental Exposures in Relation to Housing Conditions in Germany. *Environmental Research, 101*(2), 246–255. doi: 10.1016/j.envres.2005.09.008

Koopmans, R. (2010). Trade-Offs Between Equality and Difference: Immigrant Integration, Multiculturalism and the Welfare State in Cross-National Perspective. *Journal of Ethnic and Migration Studies, 36*(1), 1–26. doi: 10.1080/13691830903250881

Koopmans, R., & Veit, S. (2014). Cooperation in Ethnically Diverse Neighborhoods: A

Lost-Letter Experiment. *Political Psychology*, *35*(3), 379–400. doi: 10.1111/pops.12037

Kouros, C. D., & Cummings, E. M. (2010, February). Longitudinal Associations Between Husbands' and Wives' Depressive Symptoms. *Journal of Marriage and Family*, *72*(1), 135–147. doi: 10.1111/j.1741-3737.2009.00688.x

Kroll, M., & Schnell, R. (2016, December). Anonymisation of Geographical Distance Matrices Via Lipschitz Embedding. *International Journal of Health Geographics*, *15:1*, 1–14. doi: 10.1186/s12942-015-0031-7

Kruger, D. J., Reischl, T. M., & Gee, G. C. (2007, December). Neighborhood Social Conditions Mediate the Association Between Physical Deterioration and Mental Health. *American Journal of Community Psychology*, *40*(3-4), 261–271. doi: 10.1007/s10464-007-9139-7

Kwan, M.-P. (2012, September). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*, *102*(5), 958–968. doi: 10.1080/00045608.2012.687349

Lancee, B., & Schaeffer, M. (2015). Moving to Diversity. Residential Mobility, Changes in Ethnic Diversity, and Concerns about Immigration. In R. Koopmans, B. Lancee, & M. Schaeffer (Eds.), *Social Cohesion and Immigration in Europe and North America: Mechanisms, Conditions, and Causality* (pp. 38–55). London: Routledge.

Latkin, C. A., & Curry, A. D. (2003, March). Stressful Neighborhoods and Depression: A Prospective Study of the Impact of Neighborhood Disorder. *Journal of Health and Social Behavior*, *44*(1), 34–44. doi: 10.2307/1519814

Legewie, J., & Schaeffer, M. (2016, July). Contested Boundaries: Explaining Where Ethnoracial Diversity Provokes Neighborhood Conflict. *American Journal of Sociology*, *122*(1), 125–161. doi: 10.1086/686942

Leibniz Institute of Ecological Urban and Regional Development. (2018). *Soil Sealing. Monitor of Settlement and Open Space Development.* Retrieved October 2, 2018, from http://monitor.ioer.de/cgi-bin/wcs?MAP=S40RG_wcs.

Lersch, P. M. (2013, April). Place Stratification or Spatial Assimilation? Neighbourhood Quality Changes after Residential Mobility for Migrants in Germany. *Urban Studies*, *50*(5), 1011–1029. doi: 10.1177/0042098012464403

LeSage, J., & Pace, R. (2014, December). The Biggest Myth in Spatial Econometrics. *Econometrics*, *2*(4), 217–249. doi: 10.3390/econometrics2040217

LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics.* Boca Raton, London, New York: Chapman & Hall/CRC.

Li, C.-H. (2016, September). Confirmatory Factor Analysis with Ordinal Data: Com-

paring Robust Maximum Likelihood and Diagonally Weighted Least Squares. *Behavior Research Methods*, *48*(3), 936–949. doi: 10.3758/s13428-015-0619-7

Li, X., & Shen, C. (2013, February). Linkage of Patient Records from Disparate Sources. *Statistical Methods in Medical Research*, *22*(1), 31–38. doi: 10.1177/0962280211403600

Lim, M., Metzler, R., & Bar-Yam, Y. (2007). Global Pattern Formation and Ethnic/Cultural Violence. *Science*, *317*(5844), 1540–1544. doi: 10.1126/science.1142734

Lin, W.-F., Chen, L. H., & Li, T.-S. (2016, April). Are "We" Good? A Longitudinal Study of We-Talk and Stress Coping in Dual-Earner Couples. *Journal of Happiness Studies*, *17*(2), 757–772. doi: 10.1007/s10902-015-9621-0

Liu, H., & Reczek, C. (2012, August). Cohabitation and U.S. Adult Mortality: An Examination by Gender and Race. *Journal of Marriage and Family*, *74*(4), 794–811. doi: 10.1111/j.1741-3737.2012.00983.x

Lofors, J., & Sundquist, K. (2007, January). Low-Linking Social Capital as a Predictor of Mental Disorders: A Cohort Study of 4.5 Million Swedes. *Social Science & Medicine*, *64*(1), 21–34. doi: 10.1016/j.socscimed.2006.08.024

Maestripieri, D., M. Baran, N., Sapienza, P., & Zingales, L. (2010, September). Between- and Within-Sex Variation in Hormonal Responses to Psychological Stress in a Large Sample of College Students. *Stress*, *13*(5), 413–424. doi: 10.3109/10253891003681137

Mair, C., Diez Roux, A. V., & Galea, S. (2008, September). Are Neighborhood Characteristics Associated with Depressive Symptoms? A Review of Evidence. *Journal of Epidemiology & Community Health*, *62*(11), 940–946. doi: 10.1136/jech.2007.066605

Marques, S., & Lima, M. L. (2011, December). Living in Grey Areas: Industrial Activity and Psychological Health. *Journal of Environmental Psychology*, *31*(4), 314–322. doi: 10.1016/j.jenvp.2010.12.002

Martig, N., & Bernauer, J. (2016). Der Halo-Effekt: Diffuses Bedrohungsempfinden und SVP-Wähleranteil. *Swiss Political Science Review*, *22*(3), 385–408. doi: 10.1111/spsr.12217

McEwen, B. S. (2008, April). Central Effects of Stress Hormones in Health and Disease: Understanding the Protective and Damaging Effects of Stress and Stress Mediators. *European Journal of Pharmacology*, *583*(2-3), 174–185. doi: 10.1016/j.ejphar.2007.11.071

Meyer, R., & Bruderer Enzler, H. (2013). Geographic Information System (GIS) and Its Application in the Social Sciences Using the Example of the Swiss Environmental Survey. *Methoden, Daten, Analysen (mda)*, *7*(3), 317–346. doi: 10.12758/mda.2013

.016

Meyler, D., Stimpson, J. P., & Peek, M. K. (2007, June). Health Concordance Within Couples: A Systematic Review. *Social Science & Medicine*, *64*(11), 2297–2310. doi: 10.1016/j.socscimed.2007.02.007

Miles, R., Coutts, C., & Mohamadi, A. (2012, February). Neighborhood Urban Form, Social Environment, and Depression. *Journal of Urban Health*, *89*(1), 1–18. doi: 10.1007/s11524-011-9621-2

Morales, L., & Echazarra, A. (2013, August). Will We All Hunker Down? The Impact of Immigration and Diversity on Local Communities in Spain. *Journal of Elections, Public Opinion & Parties*, *23*(3), 343–366. doi: 10.1080/17457289.2013.808642

Müller, K., & Schwarz, C. (2017). Fanning the Flames of Hate: Social Media and Hate Crime. *SSRN Electronic Journal*, 1–76. doi: 10.2139/ssrn.3082972

Müller*, S. (2019). Räumliche Verknüpfung georeferenzierter Umfragedaten mit Geodaten: Chancen, Herausforderungen und praktische Empfehlungen. In U. Jensen, S. Netscher, & K. Weller (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten. Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 211–229). Opladen, Berlin, Toronto: Verlag Barbara Budrich.

Müller*, S., Schweers, S., & Siegers, P. (2017). *Geocoding and Spatial Linking of Survey Data—An Introduction for Social Scientists* (Tech. Rep. No. 2017|15). GESIS – Leibniz Institute for the Social Sciences.

Neumayer, E., & Plümper, T. (2016, January). W. *Political Science Research and Methods*, *4*(01), 175–193. doi: 10.1017/psrm.2014.40

Nitschke, M., Tucker, G., Simon, D. L., Hansen, A. L., & Pisaniello, D. L. (2014, January). The Link Between Noise Perception and Quality of Life in South Australia. *Noise and Health*, *16*(70), 137–142. doi: 10.4103/1463-1741.134913

Nonnenmacher, A. (2013). Zur Nachweisbarkeit von Kontexteffekten der sozialräumlichen Umgebung. In D. Oberwittler, S. Rabold, & D. Baier (Eds.), *Städtische Armutsquartiere - Kriminelle Lebenswelten?* (pp. 293–320). Wiesbaden: Springer VS. doi: 10.1007/978-3-531-93244-6_11

Nonnenmacher, A., & Friedrichs, J. (2011, February). The Missing Link: Deficits of Country-Level Studies. A Review of 22 Articles Explaining Life Satisfaction. *Social Indicators Research*, *110*, 1221–1244. doi: 10.1007/s11205-011-9981-8

Nowak, A., & Sayago-Gomez, J. (2018, May). Homeowner Preferences After September 11th, a Microdata Approach. *Regional Science and Urban Economics*, *70*, 330–351. doi: 10.1016/j.regsciurbeco.2017.10.001

Nübling, M., Andersen, H. H., & Mühlbacher, A. (2006). *Entwicklung eines Verfahrens*

zur Berechnung der körperlichen und psychischen Summenskalen auf Basis der SOEP - Version des SF 12 (Algorithmus) (Tech. Rep. No. 16). DIW Berlin, German Institute for Economic Research.

Oiamo, T. H., Baxter, J., Grgicak-Mannion, A., Xu, X., & Luginaah, I. N. (2015, September). Place Effects on Noise Annoyance: Cumulative Exposures, Odour Annoyance and Noise Sensitivity as Mediators of Environmental Context. *Atmospheric Environment, 116*, 183–193. doi: 10.1016/j.atmosenv.2015.06.024

OpenStreetMap / GEOFABRIK. (2018). *Governmental District Shapefiles.* Retrieved September 28, 2018, from https://download.geofabrik.de/europe/germany/nordrhein-westfalen/koeln-regbez-latest-free.shp.zip.

Otjes, S., & Katsanidou, A. (2017, May). Beyond Kriesiland: EU Integration as a Super Issue After the Eurocrisis. *European Journal of Political Research, 56*(2), 301–319. doi: 10.1111/1475-6765.12177

Parsons, M. A., Godøy, O., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S., & Carlson, D. (2011, December). A Conceptual Framework for Managing Very Diverse Data for Complex, Interdisciplinary Science. *Journal of Information Science, 37*(6), 555–569. doi: 10.1177/0165551511412705

Passchier-Vermeer, W., & Passchier, W. F. (2000). Noise Exposure and Public Health. *Environmental Health Perspectives, 108*(1), 123–131. doi: 10.1289/ehp.00108s1123

Pearlin, L. I., Schieman, S., Fazio, E. M., & Meersman, S. C. (2005, June). Stress, Health, and the Life Course: Some Conceptual Perspectives. *Journal of Health and Social Behavior, 46*(2), 205–219. doi: 10.1177/002214650504600206

Pedersen, E. (2015, March). City Dweller Responses to Multiple Stressors Intruding into Their Homes: Noise, Light, Odour, and Vibration. *International Journal of Environmental Research and Public Health, 12*(3), 3246–3263. doi: 10.3390/ijerph120303246

Perchoux, C., Chaix, B., Brondeel, R., & Kestens, Y. (2016, July). Residential Buffer, Perceived Neighborhood, and Individual Activity Space: New Refinements in the Definition of Exposure Areas – the RECORD Cohort Study. *Health & Place, 40*, 116–122. doi: 10.1016/j.healthplace.2016.05.004

Percivall, G. (2016). OGC's Open Standards for Geospatial Interoperability. In S. Shekhar, H. Xiong, & X. Zhou (Eds.), *Encyclopedia of GIS* (pp. 1–7). Cham: Springer International Publishing. doi: 10.1007/978-3-319-23519-6_904-2

Pescosolido, B. A. (2006, September). Of Pride and Prejudice: The Role of Sociology and Social Networks in Integrating the Health Sciences. *Journal of Health and Social Behavior, 47*(3), 189–208. doi: 10.1177/002214650604700301

Pettigrew, T. F. (1998). Intergroup Contact Theory. *Annual Review of Psychology, 49*(1),

65–85. doi: 10.1146/annurev.psych.49.1.65

Pettigrew, T. F., & Tropp, L. R. (2006). A Meta-Analytic Test of Intergroup Contact Theory. *Journal of Personality and Social Psychology*, 751–783. doi: 10.1037/0022-3514.90.5.751

Plant, R. E. (2012). *Spatial Data Analysis in Ecology and Agriculture Using R*. Boca Raton: CRC Press.

Putnam, R. D. (2007, June). E Pluribus Unum: Diversity and Community in the Twenty-first Century. The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, *30*(2), 137–174. doi: 10.1111/j.1467-9477.2007.00176.x

QGIS Development Team. (2019). *QGIS Geographic Information System*. Open Source Geospatial Foundation Project.

Quillian, L. (1995). Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe. *American Sociological Review*, 586–611. doi: 10.2307/2096296

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Raghunathan, T. (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press/-Taylor & Francis Group.

Raijman, R., Davidov, E., Schmidt, P., & Hochman, O. (2008, April). What Does a Nation Owe Non-Citizens?: National Attachments, Perception of Threat and Attitudes towards Granting Citizenship Rights in a Comparative Perspective. *International Journal of Comparative Sociology*, *49*(2-3), 195–220. doi: 10.1177/0020715208088912

Rainham, D., McDowell, I., Krewski, D., & Sawada, M. (2010, March). Conceptualizing the Healthscape: Contributions of Time Geography, Location Technologies and Spatial Ecology to Place and Health Research. *Social Science & Medicine*, *70*(5), 668–676. doi: 10.1016/j.socscimed.2009.10.035

Raju, P. L. N. (2004). Spatial Data Analysis. In M.V.K. Sivakumar, P.S. Roy, K. Harmsen, & S.K. Saha (Eds.), *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology* (pp. 151–174). Geneva, Switzerland: World Meteorological Organisation.

Rat für Sozial- und Wirtschaftsdaten — RatSWD. (2012). *Endbericht der AG „Georeferenzierung von Daten" des RatSWD — Bericht der Arbeitsgruppe und Empfehlungen des RatSWD* (Tech. Rep.).

Richardson, D. B., Volkow, N. D., Kwan, M.-P., Kaplan, R. M., Goodchild, M. F., & Croyle, R. T. (2013, March). Spatial Turn in Health Research. *Science*, *339*(6126), 1390–1392. doi: 10.1126/science.1232257

Robles, T. F., Slatcher, R. B., Trombello, J. M., & McGinn, M. M. (2014). Marital Quality and Health: A Meta-Analytic Review. *Psychological Bulletin, 140*(1), 140–187. doi: 10.1037/a0031859

Ross, C. E. (2000, June). Neighborhood Disadvantage and Adult Depression. *Journal of Health and Social Behavior, 41*(2), 177–187. doi: 10.2307/2676304

Ross, C. E., & Mirowsky, J. (2008, June). Neighborhood Socioeconomic Status and Health: Context or Composition? *City & Community, 7*(2), 163–179. doi: 10.1111/j.1540-6040.2008.00251.x

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36. doi: 10.18637/jss.v048.i02

Roswall, N., Høgh, V., Envold-Bidstrup, P., Raaschou-Nielsen, O., Ketzel, M., Overvad, K., ... Sørensen, M. (2015, March). Residential Exposure to Traffic Noise and Health-Related Quality of Life—A Population-Based Study. *PLOS ONE, 10*(3), e0120199. doi: 10.1371/journal.pone.0120199

Rüttenauer, T. (2018, February). Neighbours Matter: A Nation-Wide Small-Area Assessment of Environmental Inequality in Germany. *Social Science Research, 70*, 198–211. doi: 10.1016/j.ssresearch.2017.11.009

Rydgren, J., & Ruth, P. (2013). Contextual Explanations of Radical Right-Wing Support in Sweden: Socioeconomic Marginalization, Group Threat, and the Halo Effect. *Ethnic and Racial Studies, 36*(4), 711–728. doi: 10.1080/01419870.2011.623786

Saib, M.-S., Caudeville, J., Carre, F., Ganry, O., Trugeon, A., & Cicolella, A. (2014, April). Spatial Relationship Quantification Between Environmental, Socioeconomic and Health Data at Different Geographic Levels. *International Journal of Environmental Research and Public Health, 11*(4), 3765–3786. doi: 10.3390/ijerph110403765

Sajevaan, G. (2008). Latitude and Longitude - a Misunderstanding. *Current Science, 49*(94), 568–569.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. R. (2012, November). Linking Survey and Administrative Records: Mechanisms of Consent. *Sociological Methods & Research, 41*(4), 535–569. doi: 10.1177/0049124112460381

Sampson, R. J., Morenoff, J. D., & Gannon-Rowley, T. (2002, August). Assessing "Neighborhood Effects": Social Processes and New Directions in Research. *Annual Review of Sociology, 28*(1), 443–478. doi: 10.1146/annurev.soc.28.110601.141114

Sandberg, J. G., Harper, J. M., Jeffrey Hill, E., Miller, R. B., Yorgason, J. B., & Day, R. D. (2013). "What Happens at Home Does Not Necessarily Stay at Home": The Relationship of Observed Negative Couple Interaction With Physical Health,

Mental Health, and Work Satisfaction. *Journal of Marriage and Family*, *75*(4), 808–821. doi: 10.1111/jomf.12039

Schaeffer, M., Höhne, J., & Teney, C. (2016, February). Income Advantages of Poorly Qualified Immigrant Minorities: Why School Dropouts of Turkish Origin Earn More in Germany. *European Sociological Review*, *32*(1), 93–107. doi: 10.1093/esr/jcv091

Schlueter, E., & Scheepers, P. (2010). The Relationship Between Outgroup Size and Anti-Outgroup Attitudes: A Theoretical Synthesis and Empirical Test of Group Threat- and Intergroup Contact Theory. *Social Science Research*, 285–295. doi: 10.1016/j.ssresearch.2009.07.006

Schmidt, F. P., Basner, M., Kroger, G., Weck, S., Schnorbus, B., Muttray, A., ... Munzel, T. (2013, December). Effect of Nighttime Aircraft Noise Exposure on Endothelial Function and Stress Hormone Release in Healthy Adults. *European Heart Journal*, *34*(45), 3508–3514. doi: 10.1093/eurheartj/eht269

Schmidt-Catran, A. W., & Fairbrother, M. (2016, February). The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right. *European Sociological Review*, *32*(1), 23–38. doi: 10.1093/esr/jcv090

Schmiedeberg, C. (2015). *Regional Data in the German Family Panel (pairfam)* (Tech. Rep. No. 7).

Schneider, S. L. (2008). Anti-Immigrant Attitudes in Europe: Outgroup Size and Perceived Ethnic Threat. *European Sociological Review*, *24*(1), 53–67. doi: 10.1093/esr/jcm034

Schönwälder, K., & Söhn, J. (2009, June). Immigrant Settlement Structures in Germany: General Patterns and Urban Levels of Concentration of Major Groups. *Urban Studies*, *46*(7), 1439–1460. doi: 10.1177/0042098009104575

Schulz, M., Romppel, M., & Grande, G. (2016, December). Built Environment and Health: A Systematic Review of Studies in Germany. *Journal of Public Health*, *40*(1), 8–15. doi: 10.1093/pubmed/fdw141

Schumacker, R. E., & Lomax, R. G. (2010). *A Beginner's Guide to Structural Equation Modeling* (3rd ed ed.). New York: Routledge.

Schweers, S., Kinder-Kurlanda, K., Müller*, S., & Siegers, P. (2016, January). Conceptualizing a Spatial Data Infrastructure for the Social Sciences: An Example from Germany. *Journal of Map & Geography Libraries*, *12*(1), 100–126. doi: 10.1080/15420353.2015.1100152

Semyonov, M., Raijman, R., & Gorodzeisky, A. (2006). The Rise of Anti-Foreigner Sentiment in European Societies, 1988-2000. *American Sociological Review*, *71*(3), 426–449. doi: 10.1177/000312240607100304

Semyonov, M., Raijman, R., Tov, A. Y., & Schmidt, P. (2004, December). Population Size, Perceived Threat, and Exclusion: A Multiple-Indicators Analysis of Attitudes Toward Foreigners in Germany. *Social Science Research, 33*(4), 681–701. doi: 10.1016/j.ssresearch.2003.11.003

Serido, J., Almeida, D. M., & Wethington, E. (2004, March). Chronic Stressors and Daily Hassles: Unique and Interactive Relationships with Psychological Distress. *Journal of Health and Social Behavior, 45*(1), 17–33. doi: 10.1177/002214650404500102

Sharkey, P., & Faber, J. W. (2014). Where, When, Why, and for Whom Do Residential Contexts Matter? Moving Away from the Dichotomous Understanding of Neighborhood Effects. *Annual Review of Sociology, 40*, 559–579. doi: 10.1146/annurev-soc-071913-043350

Shepherd, D., Welch, D., Dirks, K. N., & Mathews, R. (2010, October). Exploring the Relationship Between Noise Sensitivity, Annoyance and Health-Related Quality of Life in a Sample of Adults Exposed to Environmental Noise. *International Journal of Environmental Research and Public Health, 7*(10), 3579–3594. doi: 10.3390/ijerph7103580

Shepherd, D., Welch, D., Dirks, K. N., & McBride, D. (2013, April). Do Quiet Areas Afford Greater Health-Related Quality of Life than Noisy Areas? *International Journal of Environmental Research and Public Health, 10*(4), 1284–1303. doi: 10.3390/ijerph10041284

Sheppard, E. (2005, December). Knowledge Production through Critical GIS: Genealogy and Prospects. *Cartographica: The International Journal for Geographic Information and Geovisualization, 40*(4), 5–21. doi: 10.3138/GH27-1847-QP71-7TP7

Sluiter, R., Tolsma, J., & Scheepers, P. (2015, November). At Which Geographic Scale Does Ethnic Diversity Affect Intra-Neighborhood Social Capital? *Social Science Research, 54*, 80–95. doi: 10.1016/j.ssresearch.2015.06.015

Sørensen, M., Andersen, Z. J., Nordsborg, R. B., Becker, T., Tjønneland, A., Overvad, K., & Raaschou-Nielsen, O. (2013, February). Long-Term Exposure to Road Traffic Noise and Incident Diabetes: A Cohort Study. *Environmental Health Perspectives, 121*(2), 217–222. doi: 10.1289/ehp.1205503

Spielman, S. E., & Yoo, E.-h. (2009, March). The Spatial Dimensions of Neighborhood Effects. *Social Science & Medicine, 68*(6), 1098–1105. doi: 10.1016/j.socscimed.2008.12.048

Stansfeld, S. A., Haines, M. M., Burr, M., Berry, B., & Lercher, P. (2000, January). A Review of Environmental Noise and Mental Health. *Noise and Health, 2*(8), 1–8.

Stansfeld, S. A., & Shipley, M. (2015, July). Noise Sensitivity and Future Risk of Illness and Mortality. *Science of The Total Environment, 520*, 114–119. doi:

10.1016/j.scitotenv.2015.03.053

Statistical Office of the European Union Eurostat. (2018). *World Shapefiles in EPSG:3857 and EPSG:3035.* Retrieved October 08, 2018, from https://ec.europa.eu/eurostat/cache/GISCO/distribution/v2/nuts/download/ref-nuts-2016-01m.shp.zip.

Statistical Offices of the Federation and the Länder. (2016). *Immigrant Rates. German Census 2011.* Retrieved 06 November, 2016, from https://github.com/stefmue/georefum/blob/master/data/census.attr.rda.

Statistische Ämter des Bundes und der Länder. (2011). *Zensus 2011 - Methoden und Verfahren* (Tech. Rep.).

Statistische Ämter des Bundes und der Länder. (2015). *25 Jahre Deutsche Einheit*. Wiesbaden: Statistisches Bundesamt.

Steenbeek, H. W., & van Geert, P. L. (2007, March). A Theory and Dynamic Model of Dyadic Interaction: Concerns, Appraisals, and Contagiousness in a Developmental Context. *Developmental Review, 27*(1), 1–40. doi: 10.1016/j.dr.2006.06.002

Stefanidis, A., Crooks, A., & Radzikowski, J. (2013, April). Harvesting Ambient Geospatial Information from Social Media Feeds. *GeoJournal*, *78*(2), 319–338. doi: 10.1007/s10708-011-9438-2

Stephan, W. G., Oscar Ybarra, & Morrison, K. R. (2009). Intergroup Threat Theory. In T. Nelson (Ed.), *Handbook of Prejudice, Stereotyping, and Discrimination* (pp. 43–60). New Jersey: Psychology Press.

Stimson, R. (2014). A Spatially Integrated Approach to Social Science Research. In Stimson, Robert (Ed.), *Handbook of Research Methods and Applications in Spatially Integrated Social Science* (pp. 13–25). Cheltenham, United Kingdom: Edward Elgar.

Strobl, C. (2017). Dimensionally Extended Nine-Intersection Model (DE-9IM). In S. Shekhar, Xiong, Hui, & Zhou, Xun (Eds.), *Encyclopedia of GIS* (pp. 470–476). New York, NY: Springer New York.

Sutton, T., Dassau, O., & Sutton, M. (2009). *A Gentle Introduction to GIS. Brought to you with Quantum GIS, a Free and Open SourceSoftware GIS Application for everyone.* (Tech. Rep.).

Taylor, L. K., Irvine, K., Iannotti, R., Harchak, T., & Lim, K. (2014, December). Optimal Strategy for Linkage of Datasets Containing a Statistical Linkage Key and Datasets with Full Personal Identifiers. *BMC Medical Informatics and Decision Making, 14:85*, 1–8. doi: 10.1186/1472-6947-14-85

Teney, C. (2012). Space Matters. The Group Threat Hypothesis Revisited with Geographically Weighted Regression. The Case of the NPD 2009 Electoral Success. *Zeitschrift für Soziologie, 41*(3), 207–226. doi: 10.1515/zfsoz-2012-0304

Termorshuizen, F., Braam, A. W., & van Ameijden, E. J. C. (2015, June). Neighborhood Ethnic Density and Suicide Risk Among Different Migrant Groups in the Four Big Cities in the Netherlands. *Social Psychiatry and Psychiatric Epidemiology*, *50*(6), 951–962. doi: 10.1007/s00127-014-0993-y

Thoits, P. A. (1994, June). Stressors and Problem-Solving: The Individual as Psychological Activist. *Journal of Health and Social Behavior*, *35*(2), 143–160. doi: 10.2307/2137362

Thoits, P. A. (2010, October). Stress and Health: Major Findings and Policy Implications. *Journal of Health and Social Behavior*, *51*(1 Suppl), S41–S53. doi: 10.1177/0022146510383499

Thoits, P. A. (2011, June). Mechanisms Linking Social Ties and Support to Physical and Mental Health. *Journal of Health and Social Behavior*, *52*(2), 145–161. doi: 10.1177/0022146510395592

Thompson, C. W., Roe, J., Aspinall, P., Mitchell, R., Clow, A., & Miller, D. (2012, April). More Green Space Is Linked to Less Stress in Deprived Communities: Evidence from Salivary Cortisol Patterns. *Landscape and Urban Planning*, *105*(3), 221–229. doi: 10.1016/j.landurbplan.2011.12.015

Thorlindsson, T., Valdimarsdottir, M., & Hrafn Jonsson, S. (2012, July). Community Social Structure, Social Capital and Adolescent Smoking: A Multi-Level Analysis. *Health & Place*, *18*(4), 796–804. doi: 10.1016/j.healthplace.2012.03.013

Timmermans, S., & Haas, S. (2008, July). Towards a Sociology of Disease. *Sociology of Health & Illness*, *30*(5), 659–676. doi: 10.1111/j.1467-9566.2008.01097.x

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240. doi: 10.2307/143141

Tolsma, J., & van der Meer, T. W. G. (2017, March). Losing Wallets, Retaining Trust? The Relationship Between Ethnic Heterogeneity and Trusting Coethnic and Non-coethnic Neighbours and Non-neighbours to Return a Lost Wallet. *Social Indicators Research*, *131*(2), 631–658. doi: 10.1007/s11205-016-1264-y

van de Schoot, R., Lugtig, P., & Hox, J. (2012, July). A Checklist for Testing Measurement Invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. doi: 10.1080/17405629.2012.686740

Wang, Q., Phillips, N. E., Small, M. L., & Sampson, R. J. (2018, July). Urban Mobility and Neighborhood Isolation in America's 50 Largest Cities. *Proceedings of the National Academy of Sciences*, *115*(30), 7735–7740. doi: 10.1073/pnas.1802537115

Weins, C. (2011). Gruppenbedrohung oder Kontakt? *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *63*, 481–499. doi: 10.1007/s11577-011-0141-6

Weßling, K. (2016). *The Influence of Socio-Spatial Contexts on Transitions from School*

to Vocational and Academic Training in Germany (Unpublished doctoral dissertation). Eberhard Karls Universität, Tübingen.

Weßling, K., Hartung, A., & Hillmert, S. (2015). Spatial Structure Counts: The Relevance of Regional Labour-Market Conditions for Educational Transitions to Vocational Training. *Empirical Research in Vocational Education and Training, 7:12,* 1–20. doi: 10.1186/s40461\%0011015\%00110024\%00116

Wimmer, A., & Lewis, K. (2010, September). Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook. *American Journal of Sociology, 116*(2), 583–642. doi: 10.1086/653658

Wolch, J. R., Byrne, J., & Newell, J. P. (2014, May). Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landscape and Urban Planning, 125,* 234–244. doi: 10.1016/j.landurbplan.2014.01.017

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach* (5th ed.). Mason: South Western.

World Health Organization. (2016). *Urban Green Spaces and Health: A Review of Evidence*. Copenhagen: WHO Regional Office for Europe.

Wu, J. (2007). Scale and Scaling: A Cross-Disciplinary Perspective. In J. Wu & R. J. Hobbs (Eds.), *Key Topics in Landscape Ecology* (pp. 115–142). Cambridge: Cambridge University Press. (OCLC: ocm70671785)

You, L., Spoor, M., Ulimwengu, J., & Zhang, S. (2011, December). Land Use Change and Environmental Stress of Wheat, Rice and Corn Production in China. *China Economic Review, 22*(4), 461–473. doi: 10.1016/j.chieco.2010.12.001

Zandbergen, P. A. (2014). Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data. *Advances in Medicine, 2014,* 1–14. doi: 10.1155/2014/567049

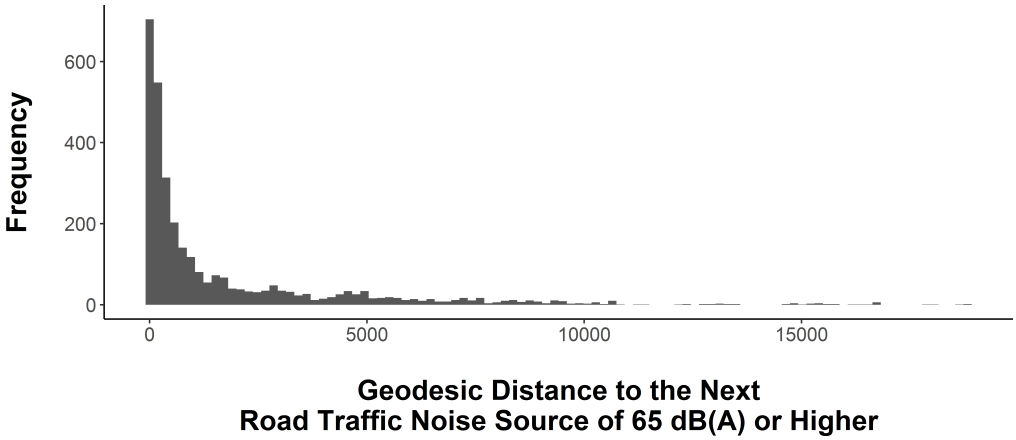
Zhang, L.-C. (2012, February). Topics of Statistical Theory for Register-Based Statistics and Data Integration: Developing Theory for Data Integration. *Statistica Neerlandica, 66*(1), 41–63. doi: 10.1111/j.1467-9574.2011.00508.x

Zhong, Z.-N., Jing, N., Chen, L., & Wu, Q.-Y. (2004, May). Representing Topological Relationships Among Heterogeneous Geometry-Collection Features. *Journal of Computer Science and Technology, 19*(3), 280–289. doi: 10.1007/BF02944898

Zwickl, K., Ash, M., & Boyce, J. K. (2014, November). Regional Variation in Environmental Inequality: Industrial Air Toxics Exposure in U.S. Cities. *Ecological Economics, 107,* 494–509. doi: 10.1016/j.ecolecon.2014.09.013

* *Author's name before marriage.*

Appendix



Data Source: Georeferenced German General Social Survey 2014 (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Figure A.1: Distribution of Geodesic Distances to the Next Road Traffic Noise Source of ≥ 65 dB(A) (N = 3,163)

Table A.1: SF-12 Items in the Georeferenced GGSS

Item	Variable Name	Question Text (German)	Valid Answer Categories
Physical Component Score			
General Health			
	V225	"How would you describe your health in general?"	"Very good", "Good", "Satisfactory", "Poor", "Bad"
	V226	"How would you describe your health in general?"	"Excellent", "Very good", "Good", "Satisfactory", "Poor", "Bad"
Physical Functioning			
	V227	"When you climb stairs, i.e. go up several floors on foot: Does your state of health affect you greatly, slightly or not at all?"	"Greatly", "Slightly", "Not at all"
	V228	"And what about having to cope with other tiring everyday tasks, e.g. lifting something heavy or performing tasks requiring agility (...)"	"Greatly", "Slightly", "Not at all"
Role Physical			
	V235	"And how often did it occur during the past four weeks (...) that, due to PHYSICAL HEALTH PROBLEMS, you accomplished less than you wanted to at work or in everyday tasks?"	"Always", "Often", "Sometimes", "Almost never", "Never"
	V236	"(...) that you were therefore limited in the kind of activities you carried out?"	"Always", "Often", "Sometimes", "Almost never", "Never"
Bodily Pain			
	V233	"During the past four weeks, how often did it occur (...) that you had strong physical pains?"	"Always", "Often", "Sometimes", "Almost never", "Never"
Mental Component Score			
Vitality			
	V232	"(...) that you felt a lot of energy?"	"Always", "Often", "Sometimes", "Almost never", "Never"
Social Functioning			
	V239	"(...) that, due to PHYSICAL OR MENTAL HEALTH, your scope of social activities was restricted, i.e. social contacts with friends, acquaintances or relatives?"	"Always", "Often", "Sometimes", "Almost never", "Never"
Role Emotional			
	V237	" (...) that, due to MENTAL HEALTH OR EMOTIONAL PROBLEMS, you accomplished less than you wanted to at work or in everyday tasks?"	"Always", "Often", "Sometimes", "Almost never", "Never"
	V238	"(...) that you were therefore limited in the kind of activities you carried out?"	"Always", "Often", "Sometimes", "Almost never", "Never"
Mental Health			
	V229	"(...) that you felt rushed or pressed for time?"	"Always", "Often", "Sometimes", "Almost never", "Never"
	V230	"(...) that you felt downhearted and blue?"	"Always", "Often", "Sometimes", "Almost never", "Never"

Table A.2: Test for Metric Invariance Between Unmarried and Married People (N Unmarried = 1383; N Married = 1755)

	χ^2	df	p	CFI	TLI	RMSEA	BIC	AIC
Free	1761.817	169	0	.901	.865	.049	143220.6	142609.5
Fixed	1733.534	162	0	.904	.864	.049	143248.7	142595.2

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Table A.3: Standardized Linear Regression Coefficients of the SEM Model with Control Variables (N Unmarried = 1383; N Married = 1755)

	Unmarried			Married		
	β	SE	95% CI	β	SE	95% CI
Physical Component Score						
Noise Annoyance	-.114***	.026	[-.165, -.062]	-.098***	.027	[-.151, -.044]
Road Traffic Noise	-.043	.127	[-.291, .205]	.074	.141	[-.203, .351]
Age	-.521***	.028	[-.576, -.467]	-.404***	.028	[-.459, -.349]
Gender	-.054*	.024	[-.101, -.007]	-.023	.027	[-.077, .030]
Income	.062*	.026	[.012, .112]	.095**	.033	[.031, .159]
Education	.175***	.024	[.129, .222]	.156***	.022	[.113, .199]
Smoking	-.076*	.032	[-.138, -.013]	-.061†	.032	[-.124, .001]
Municipality Size	.024	.124	[-.220, .267]	-.038	.140	[-.313, .237]
Mental Component Score						
Noise Annoyance	-.134***	.032	[-.197, -.071]	-.090**	.031	[-.151, -.030]
Road Traffic Noise	-.339†	.181	[-.693, .015]	.195	.155	[-.109, .500]
Age	-.193***	.036	[-.264, -.122]	-.050	.033	[-.115, .015]
Gender	-.095**	.030	[-.153, -.036]	-.054†	.030	[-.112, .005]
Income	.039	.032	[-.025, .103]	.053	.040	[-.026, .132]
Education	.065*	.033	[.000, .129]	.095***	.028	[.039, .150]
Smoking	-.057	.038	[-.133, .018]	-.027	.034	[-.095, .041]
Municipality Size	.304†	.174	[-.037, .646]	-.099	.145	[-.383, .185]
Noise Annoyance						
Road Traffic Noise	.335†	.176	[-.011, .681]	.649***	.140	[.375, .924]
Age	.057	.038	[-.018, .132]	.109***	.030	[.051, .168]
Gender	.072*	.030	[.013, .131]	.021	.024	[-.025, .068]
Income	-.017	.032	[-.079, .045]	-.066*	.029	[-.122, -.009]
Education	.003	.032	[-.060, .066]	.039	.030	[-.020, .098]
Smoking	-.016	.037	[-.089, .056]	-.017	.034	[-.082, .049]
Municipality Size	-.195	.176	[-.540, .150]	-.525***	.138	[-.796, -.255]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Table B.1: Loadings of the Threat Variables on 2 PCA Components and 1 PCA Component (N = 1,192)

	2 Components		1 Component
	PC1 (.398%)	PC1 (.328%)	PC1 (.566%)
Increase crime	.86	.08	.71
Good for economy	.14	.90	.69
Take away jobs	.73	.35	.79
New ideas and cultures	.37	.77	.77
Undermine culture	.75	.35	.80

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Table B.2: Standardized Coefficients of a Simultaneous Estimated Spatial Lag Y Regression Model for the Default Halo Operationalization (N = 744)

	β	SE	CI
Halo	-0.032	0.030	[-0.091, 0.03]
Threat ρ	0.297*	0.122	[0.058, 0.54]
Inhabitants	-0.132**	0.046	[-0.221, -0.04]
Flat Size	0.002	0.040	[-0.076, 0.08]
Municipality Size			
Rural Community (*Ref.*)			
Small Town	0.099	0.097	[-0.090, 0.29]
City	-0.011	0.103	[-0.213, 0.19]
Big City	-0.026	0.120	[-0.262, 0.21]
Age	0.043	0.034	[-0.024, 0.11]
Gender (female)	-0.068	0.071	[-0.208, 0.07]
Education			
Low (*Ref.*)			
Medium	-0.205†	0.124	[-0.448, 0.04]
Advanced	-0.657***	0.142	[-0.935, -0.38]
High	-0.998***	0.145	[-1.282, -0.71]
Income	-0.010	0.041	[-0.089, 0.07]
Unemployment	0.306†	0.161	[-0.010, 0.62]
Eastern Germany	0.195*	0.082	[0.035, 0.36]
Homeownership	-0.060	0.075	[-0.207, 0.09]
Intercept	0.441**	0.148	[0.150, 0.73]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced German General Social Survey (GESIS - Leibniz Institute for the Social Sciences, 2015, 2018)

Table C.1: Unimputed Standardized Regression Coefficients for the Baseline Model Between Increasing Geographic Sizes of Soil Sealing, (N = 2,481; Clustered Standard Errors)

	100m × 100m			500m Buffer			1000m Buffer			2000m Buffer		
	β	SE	CI	β	SE	CI	β	SE	CI	β	SE	CI
Intercept	.913***	.224	[.471, 1.354]	-.435*	.194	[-.817, -.052]	-1.115***	.192	[-1.493, -.737]	-1.758***	.194	[-2.140, -1.376]
Age	-.002	.003	[-.007, .003]	.002	.002	[-.002, .006]	.004†	.002	[.000, .008]	.004*	.002	[.000, .008]
Gender (Female)	.061	.050	[-.038, .160]	.060	.049	[-.036, .157]	.065	.050	[-.033, .163]	.082†	.046	[-.008, .172]
Education												
Low (Ref.)												
Medium	.079	.088	[-.094, .251]	.033	.077	[-.118, .184]	.025	.084	[-.140, .190]	.045	.083	[-.118, .208]
High	.049	.081	[-.111, .208]	.189**	.072	[.048, .330]	.293***	.078	[.139, .447]	.379***	.081	[.221, .538]
Homeownership	-.501***	.075	[-.650, -.353]	-.548***	.063	[-.672, -.425]	-.574***	.064	[-.700, -.448]	-.565***	.064	[-.690, -.440]
Household Size	-.036	.034	[-.103, .030]	-.068*	.028	[-.123, -.012]	-.065*	.027	[-.118, -.012]	-.059*	.024	[-.106, -.011]
Inhabitants	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]
Income	-.027*	.013	[-.053, -.001]	-.008	.011	[-.031, .014]	-.007	.012	[-.031, .017]	-.005	.012	[-.029, .020]
Migrant	.166	.102	[-.035, .368]	.348***	.088	[.175, .521]	.456***	.087	[.285, .627]	.424***	.086	[.254, .593]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017)

Table C.2: Standardized Regression Coefficients for the Interaction Model Between Increasing Geographic Sizes of Soil Sealing (N = 3,852; Clustered Standard Errors)

	100m × 100m			500m Buffer			1000m Buffer			2000m Buffer		
	β	SE	CI	β	SE	CI	β	SE	CI	β	SE	CI
Intercept	1.026***	.167	[.698, 1.354]	-.230†	.128	[-.481, .020]	-.973***	.151	[-1.270, -.676]	-1.566***	.154	[-1.868, -1.265]
Age	-.005*	.002	[-.008, -.001]	.000	.001	[-.003, .002]	.002	.002	[-.001, .005]	.002	.002	[-.001, .005]
Gender (Female)	.078†	.045	[-.010, .167]	.065†	.034	[.000, .131]	.088*	.041	[.009, .168]	.090*	.038	[.015, .164]
Education												
Low (Ref.)												
Medium	-.043	.070	[-.181, .095]	-.006	.057	[-.117, .105]	-.003	.070	[-.140, .135]	.009	.072	[-.133, .151]
High	-.067	.067	[-.199, .065]	.138**	.054	[.032, .243]	.275***	.069	[.140, .409]	.337***	.070	[.199, .474]
Homeownership	-.424***	.066	[-.554, -.294]	-.453***	.049	[-.549, -.356]	-.543***	.057	[-.654, -.432]	-.545***	.057	[-.656, -.433]
Household Size	-.091***	.022	[-.134, -.047]	-.089***	.016	[-.121, -.058]	-.105***	.019	[-.143, -.067]	-.098***	.019	[-.135, -.060]
Number of Inhabitants	.000**	.000	[.000, .000]	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]	.000***	.000	[.000, .000]
Income	-.009	.011	[-.031, .012]	-.003	.009	[-.020, .014]	-.001	.010	[-.022, .019]	-.001	.011	[-.023, .020]
Migrant	.681*	.327	[.031, 1.332]	.567*	.220	[.134, 1.000]	.691**	.238	[.222, 1.160]	.568*	.225	[.127, 1.010]
Income × Migrant	-.046	.029	[-.104, .012]	-.021	.019	[-.058, .016]	-.021	.020	[-.060, .019]	-.014	.019	[-.051, .023]

† $p \leq .1$; * $p \leq .05$; ** $p \leq .01$; *** $p \leq .001$

Data Source: Georeferenced GESIS Panel 2014 (GESIS - Leibniz Institute for the Social Sciences, 2017)

This book demonstrates the use of georeferenced data for social science survey research which builds upon survey data enriched with geo-coordinates. It reviews the prerequisites and challenges of applying these data to different social science research questions, highlighting the different branches of an interdisciplinary effort. At the center of this presentation is the method of spatial linking: the combination of georeferenced survey data with information from auxiliary geospatial data sources. A collection of spatial linking methods is applied in this book's empirical applications which underline these methods' flexibility in different social science sub-disciplines, such as health and family, political attitudes, and environmental inequalities. For this purpose, georeferenced survey data from the German General Social Survey (GGSS) 2014 and the GESIS Panel are used. These empirical applications are part of an emerging field of research for social scientists, requiring new analytic skills from diverse and foreign disciplines, like ecology and engineering. Navigating the organizational and technical requirements for the analysis of georeferenced survey data enables researchers to answer new and innovative research questions.

Dieses Buch beschäftigt sich mit der Nutzung georeferenzierter Daten in der sozialwissenschaftlichen Umfrageforschung, deren Ausgangspunkt Umfragedaten sind, die mit Geokoordinaten angereichert wurden. Es widmet sich den Voraussetzungen und Herausforderungen, solche Daten für verschiedene sozialwissenschaftliche Fragestellungen nutzbar zu machen und betont dabei die verschiedenen interdisziplinären Verzweigungen dieses Unterfangens. Im Mittelpunkt der Präsentation steht die Methode der räumlichen Verknüpfung: die Kombination georeferenzierter Umfragedaten mit Informationen aus externen Geodatenquellen. Anhand mehrerer, aus unterschiedlichen Subdisziplinen der Sozialwissenschaften stammender empirischer Anwendungen im Bereich Familie und Gesundheit, politische Einstellungen sowie Umwelt und Ungleichheit wird die Flexibilität der Methode in Form verschiedener räumlicher Verknüpfungen betont. Dazu werden georeferenzierte Umfragedaten der Allgemeinen Bevölkerungsumfrage Sozialwissenschaften (ALLBUS) 2014 und dem GESIS Panel 2014 verwendet. Diese empirischen Anwendungen sind Teil eines aufstrebenden Forschungsfelds für Sozialforschende, welches neue analytische Fertigkeiten aus verschiedenen anderen Fachbereichen wie der Ökologie oder des Ingenieurswesen erfordert. Werden die organisatorischen und technischen Anforderungen zur Analyse georeferenzierter Umfragedaten gemeistert, eröffnet sich Forschenden die Möglichkeit, neue und innovative Fragestellungen zu beantworten.

gesis Leibniz-Institut für Sozialwissenschaften

ISBN 978-3-86819-040-3 (print)
ISBN 978-3-86819-039-7 (eBook)
ISSN 1869-2869
15,- EUR